Einfluß von Fehlern in den Ausgangsdaten eines linearen Gleichungssystems auf die Lösung

Inaugural - Dissertation

zur Erlangung der Doktorwürde
der Hohen Naturwissenschaftlichen Fakultät
der Julius-Maximilians-Universität
Würzburg

von

Horst Mayer
aus Berlin

Würzburg 1970

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41543520_0019

Einfluß von Fehlern in den Ausgangsdaten eines linearen Gleichungssystems auf die Lösung

Inaugural - Dissertation

zur Erlangung der Doktorwürde
der Hohen Naturwissenschaftlichen Fakultät
der Julius-Maximilians-Universität
Würzburg

von
Horst Mayer
aus Berlin

Würzburg 1970

Referent:	Prof. Dr. P.O. Runck
Korreferent:	Prof. Dr. W. Velte
Eingereicht am:	15. Dez. 1969
Mündliche Prüfung:	11. Feb. 1970

UNIV.
LEIDEN
BIBL.

Ich versichere an Eides Statt, daß die gemachten Angaben der Wahrheit entsprechen und daß ich außer den angegebenen keine weiteren Hilfsmittel benutzt habe.

Die vorliegende Arbeit entstand am Institut für Angewandte Mathematik der Bayerischen Julius - Maximilians - Universität in Würzburg unter der Leitung von Herrn Prof.Dr.P.O.Runck. Ihm danke ich recht herzlich für die Betreuung, die ich während der Anfertigung der Arbeit erfuhr.

Den Herren Dietrich, Löhe und Rausche danke ich für ihre Hilfe bei der Anfertigung der Algolprogramme.

Würzburg, im Dezember 1969

Horst Mayer

I N H A L T

	Seite
Einleitung	1
I. Kapitel : Vorbereitungen	
I.1 Grundlagen aus der Matrizenlehre	3
I.2 Grundlagen aus der Wahrscheinlichkeitstheorie	9
II. Kapitel : Untersuchungen über den Einfluß von Fehlern in den Ausgangsdaten eines linearen Gleichungssystems auf die Lösung	
II.1. Abschätzungen für den maximalen Defektvektor bei Ungenauigkeiten in den Ausgangsdaten	14
II.2. Abhängigkeit der Lösung eines linearen Gleichungssystems von zufälligen Fehlern in den Ausgangsdaten	32
II.3. Experimentelle Untersuchungen für zufällige Fehler in den Ausgangsdaten	39
II.4. Stabilität von Matrizen. Konditionszahlen	44
III. Kapitel : Numerische Auswertung der Ergebnisse	
III.1. Numerische Auswertung der Ergebnisse bei bekannter inverser Matrix A^{-1}	49
III.2. Numerische Auswertung der Ergebnisse ohne Kenntnis der inversen Matrix A^{-1}	56
IV. Kapitel : Algolprogramme	
IV.1. Allgemeine Bemerkungen	81
IV.2. Algolprogramme mit Diskussion	81
Literaturverzeichnis	107

EINLEITUNG :

In der vorliegenden Arbeit wird untersucht, wie sich bei einem linearen Gleichungssystem $Ax = b$ Ungenauigkeiten in den Ausgangsdaten auf die Lösung des linearen Gleichungssystems auswirken. Die Art der Fehler kann sehr unterschiedlich sein. Wir werden uns zunächst mit den beiden folgenden Fehlertypen befassen:

Numerisch gegeben sei das lineare Gleichungssystem

$$Ax = b \; ; \; \det A \neq 0 \quad x \text{ Lösung}$$

Wir betrachten die benachbarten Gleichungssysteme :

I. $(A+R)(x+\delta x) = b+f \quad ; \; R = (r_{ij}) \text{ mit } |r_{ij}| \leq \varepsilon_{ij} \; ; \; \det(A+R) \neq 0$

$$f = (f_i) \text{ mit } |f_i| \leq \delta_i$$

II. $(A+R)(x+\delta x) = b+f \quad ; \; \|R\| \leq \varepsilon \; , \; \|f\| \leq \delta \; , \; \det(A+R) \neq 0$

Vektor- und Matrixnorm seien verträglich.

Von wesentlicher Bedeutung ist nun die folgende Größe :

$$\rho := \max_{R,f} \|\delta x\|$$

ρ gibt an, wie stark sich innerhalb der gegebenen Fehlergrenzen die Lösung eines benachbarten Gleichungssystems von der Lösung des numerisch gegebenen Gleichungssystems bezüglich einer Vektornorm $\| \; \|$ maximal unterscheiden kann.

Der Abschätzung dieser Größe ρ gilt ein wesentlicher Teil der Arbeit. Für Spezialfälle hat sich bereits 1934 Wittmeyer [8] in seiner Dissertation mit diesem Problem beschäftigt. Später haben vor allem Collatz [1] und Dück [2] Beiträge hierzu geliefert.

Ausführlich beschäftigen wir uns auch mit der Auswertung dieser Abschätzungen. In die Abschätzungen geht die inverse Matrix A^{-1} explizit ein. Bei Kenntnis der inversen Matrix ist die numerische Auswertung im allgemeinen ohne großen Aufwand möglich. Es werden Verfahren angegeben, mit denen man auch ohne Kenntnis der inversen Matrix die Abschätzungen auswerten kann.

In vielen Fällen kann man davon ausgehen, daß die Fehler in den Ausgangsdaten eines linearen Gleichungssystems zufällig sind. In dieser Arbeit werden auch Untersuchungen angestellt über den Einfluß von zufälligen Fehlern in den Ausgangsdaten auf die Lösung des linearen Gleichungssystems.

Wir interessieren uns bezüglich der zufälligen Variablen, die durch die Norm des Defektvektors δx bestimmt ist, zu vorgegebener Sicherheitswahrscheinlichkeit p für den Quantil γ der Ordnung p.

$$W\left(\left\{ \|\delta x\| \leq \gamma \right\} \right) \geq p$$

Es ist gelungen unter speziellen Voraussetzungen über die Art der Fehler und über die Struktur der Koeffizientenmatrix Abschätzungen für die Größe γ anzugeben, die sich numerisch auswerten lassen.

Es wurden ferner eine Reihe von Algolprogrammen angefertigt, mit deren Hilfe ein Teil der Ergebnisse dieser Arbeit getestet wurde.

I. KAPITEL.

Vorbereitungen

§ 1. Grundlagen aus der Matrizenlehre

In diesem Paragraphen werden einige Sätze und Begriffe aus der Matrizenlehre zusammengestellt, die in dieser Arbeit von Bedeutung sind. (vgl. Zurmühl [9])

Satz I.1 :

Eine positiv definite (positiv semidefinite) Matrix A hat nur positive (nicht negative) Eigenwerte. Sind λ_1, λ_n ein kleinster bzw. größter Eigenwert von A, so gilt die folgende Abschätzung:

$$\lambda_1 \bar{x}^T x \leq \bar{x}^T A x \leq \lambda_n \bar{x}^T x$$

Das „=“-Zeichen tritt genau dann ein, wenn x ein zu λ_1 bzw. λ_n gehörender Eigenvektor ist.

Definition:

Die Eigenwerte $\Lambda_1, \ldots, \Lambda_n$ von $\bar{A}^T A$ heißen die singulären Werte von A.

Satz I.2 :

A sei eine beliebige Matrix, dann ist $\bar{A}^T A$ stets positiv semidefinit. $\bar{A}^T A$ ist genau dann positiv definit, wenn det $A \neq 0$. Es gilt die Abschätzung: $\Lambda_1 \leq |\lambda_i|^2 \leq \Lambda_n \quad i = 1,2,\ldots,n$. Ist A eine hermitesche Matrix, so gilt: $\Lambda_n = |\lambda_n|^2$

Satz I.3 :

$\lambda_1, \lambda_2, \ldots, \lambda_n$ seien die Eigenwerte einer Matrix A. Dann gilt:

I. $\lambda_i \neq 0 \quad i = 1,2,\ldots,n$ gilt genau dann, wenn det $A \neq 0$

II. Wenn det $A \neq 0$, dann sind $\lambda_i^{-1} \quad i = 1,2,\ldots,n$ die Eigenwerte von A^{-1}.

III. $\bar{A}^T A$ und $A \bar{A}^T$ haben stets die gleichen Eigenwerte.

Für eine Reihe von Untersuchungen ist es von Bedeutung in der Menge der Vektoren und Matrizen einen Abstandsbegriff zur Verfügung zu haben. Die folgenden Ausführungen gelten der Einführung und

Diskussion von Vektor- und Matrixnormen.

Definition :

$\| \ \| : \left.\begin{matrix} \mathbb{R}^n \\ \mathbb{C}^n \end{matrix}\right\} \longrightarrow \mathbb{R}$ heisst Vektornorm, wenn folgende Eigenschaften erfüllt sind:

1. $\|x\| \geq 0$ „=" genau dann, wenn $x = v$
2. $\|x+y\| \leq \|x\| + \|y\|$
3. $\|\alpha x\| = |\alpha| \|x\| \quad \alpha \in \mathbb{R}, \mathbb{C}$

Bei der Einführung einer Matrixnorm erweist es sich als zweckmäßig, noch eine Zusatzforderung zu stellen.

Definition :

$\| \ \| : \left\{ A = (a_{ij}) \mid a_{ij} \in \mathbb{R}, \mathbb{C} \right\} \longrightarrow \mathbb{R}$

heißt multiplikative Matrixnorm, wenn folgende Eigenschaften erfüllt sind :

1. $\|A\| \geq 0$ „=" genau dann, wenn $A = 0$
2. $\|A+B\| \leq \|A\| + \|B\|$
3. $\|\alpha A\| = |\alpha| \|A\| \quad \alpha \in \mathbb{R}, \mathbb{C}$
4. $\|AB\| \leq \|A\| \ \|B\|$

Satz I.4 :

Der Raum der Vektoren versehen mit einer Vektornorm und der Raum der Matrizen versehen mit einer Matrixnorm bilden einen Banachraum, d.h. einen linearen normierten vollständigen Raum.

Bezeichnung :

M_n sei der Banachraum der (n,n)-Matrizen.

V_n sei der Banachraum der n-dimensonalen Vektoren.

Zwischen Matrizen und Vektoren ist eine Verknüpfung definiert.
Es erweist sich als zweckmäßig zwischen beiden Normen eine Verträglichkeitsbedingung zu fordern.

Definition :

Eine Vektornorm $\|x\|$ und eine Matrixnorm $\|A\|$ heißen miteinander verträglich, wenn für eine beliebige Matrix A und einen beliebigen Vektor x gilt : $\|Ax\| \leq \|A\| \, \|x\|$

Wir wollen uns nun einen Überblick über verträgliche Matrix- und Vektornormen verschaffen und einige wichtige Beispiele angeben.

Definition :

$\|x\|$ sei eine gegebene Vektornorm. $c \geq 0$ heißt eine Schranke von A, wenn für einen beliebigen Vektor x gilt : $\|Ax\| \leq c\|x\|$

Bemerkung :

Die mit x verträglichen Matrixnormen sind Schranken.

Definition :

Unter der kleinsten oberen Schranke lub(A) versteht man :

$$\operatorname{lub}(A) := \sup_{x \neq \vartheta} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

Satz I.5 :

lub(A) hat die Eigenschaften einer multiplikativen Matrixnorm.

Beweis :

Zu 1.

Es gilt : $\max_{\|x\|=1} \|Ax\| \geq 0$ für alle A und x

Zu $A \neq 0$ existiert stets ein Vektor x mit $Ax \neq \vartheta$

Daraus folgt : $\operatorname{lub}(A) = 0$ genau dann, wenn $A = 0$

Zu 2.

Es gilt : $\operatorname{lub}(A+B) = \max_{\|x\|=1} \|(A+B)x\| = \max_{\|x\|=1} \|Ax + Bx\| =$

$$\leq \max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\| = \operatorname{lub}(A) + \operatorname{lub}(B)$$

Zu 3.

Es gilt : $\operatorname{lub}(\alpha A) = \max_{\|x\|=1} \|\alpha Ax\| = |\alpha| \max_{\|x\|=1} \|Ax\| = |\alpha| \operatorname{lub}(A)$

Zu 4.

Es gilt : $\mathrm{lub}(AB) = \sup_{x \neq \mathfrak{o}} \frac{\|ABx\|}{\|x\|} = \sup_{x \neq \mathfrak{o}} \left(\frac{\|ABx\|}{\|Bx\|} \frac{\|Bx\|}{\|x\|} \right) =$

$$\leq \sup_{x \neq \mathfrak{o}} \frac{\|ABx\|}{\|Bx\|} \sup_{x \neq \mathfrak{o}} \frac{\|Bx\|}{\|x\|} = \sup_{z \neq \mathfrak{o}} \frac{\|Az\|}{\|z\|} \sup_{x \neq \mathfrak{o}} \frac{\|Bx\|}{\|x\|} =$$

$$= \mathrm{lub}(A)\ \mathrm{lub}(B)$$

Folgerungen :

1. Zu jeder Vektornorm existiert mindestens eine mit ihr verträgliche multiplikative Matrixnorm - z.B. die lub - Norm. Die lub-Norm nennt man auch die der Vektornorm zugeordnete Matrixnorm.
2. Wenn $\|A\|$ irgendeine mit $\|x\|$ verträgliche Matrixnorm ist, dann gilt : $\|A\| \geq \mathrm{lub}(A)$.
3. $\|A\|$ sei eine beliebige Matrixnorm. Für jede Matrix A existiere ein Vektor $x \neq \mathfrak{o}$, so daß $\|Ax\| = \|A\|\ \|x\|$ gilt, dann ist $\|A\| = \mathrm{lub}(A)$.

Wir haben gezeigt, daß zu einer gegebenen Vektornorm stets eine verträgliche Matrixnorm existiert. Die Umkehrung dieses Sachverhalts wird in dem folgenden Satz gezeigt.

Satz I.6 :

$\|A\|$ sei eine beliebige multiplikative Matrixnorm. Dann existiert stets eine mit ihr verträgliche Vektornorm.

Beweis :

Durch die folgende Festsetzung ist eine Vektornorm definiert :

$$x = (x_i) \longleftrightarrow X = \begin{pmatrix} x_1 & 0 \cdots 0 \\ \vdots & \vdots \\ x_n & 0 \cdots 0 \end{pmatrix} \quad ; \qquad \|x\| := \|X\|$$

Wegen $\|Ax\| = \|AX\| \leq \|A\|\ \|X\| = \|A\|\ \|x\|$ sind beide Normen miteinander verträglich.

Wir wollen noch auf den Zusammenhang zwischen dem Spektralradius einer Matrix A und einer gegebenen Matrixnorm hinweisen.

Definition :

A sei eine Matrix mit den Eigenwerten $\lambda_1, \lambda_2, \ldots, \lambda_n$, dann heißt $\rho(A) := \max_{i=1,\ldots,n} |\lambda_i|$ der Spektralradius von A.

Satz I.7 :

$\|x\|$ und $\|A\|$ seien verträgliche Normen, dann gelten die Abschätzungen:

$$\rho(A) \leq \text{lub}(A) \leq \|A\|$$

Nach diesen Vorbereitungen wollen wir einige wichtige Vektor- und Matrixnormen angeben.

Vektornormen :

1. $\|x\|_1 := \max_{i=1,\ldots,n} |x_i|$ Maximumnorm
2. $\|x\|_2 := \sum_{i=1}^{n} |x_i|$ Betragssummennorm
3. $\|x\|_3 := (\bar{x}^T x)^{1/2}$ Euklidsche Vektornorm

Matrixnormen :

1. $\|A\|_1 := Z(A) := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$ Zeilennorm
2. $\|A\|_2 := S(A) := \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$ Spaltennorm
3. $\|A\|_3 := H(A) := (\Lambda_n)^{1/2}$ Spektralnorm
4. $\|A\| := N(A) := (\sum_{i,j=1}^{n} |a_{ij}|^2)^{1/2}$ Euklidsche Norm
5. $\|A\| := M(A) := n \max_{i,j=1,\ldots,n} |a_{ij}|$ Gesamtnorm

Auf den Nachweis der eben angegebenen Vektor- und Matrixnormen soll hier verzichtet werden. Die Indizes an den verschiedenen Normen sollen bei späteren Betrachtungen eine bequeme Schreibweise ermöglichen. Eine Vektornorm und die ihr zugeordnete Matrixnorm wird mit dem gleichen Index bezeichnet.

Satz I.8 :

1. Für $\|x\|_1 := \max_{i=1,\dots,n} |x_i|$ gilt : $Z(A) = \text{lub}(A)$
2. Für $\|x\|_2 := \sum_{i=1}^{n} |x_i|$ gilt : $S(A) = \text{lub}(A)$
3. Für $\|x\|_3 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$ gilt : $H(A) = \text{lub}(A)$

In der folgenden Tabelle werden die letzten Ergebnisse übersichtlich zusammengestellt.

Vektornorm	verträgliche Matrixnorm
$\|x\|_1 := \max_{i=1,\dots,n} \|x_i\|$	$\|A\|_1 := \max_{j=1,\dots,n} \sum_{i=1}^{n} \|a_{ij}\| = Z(A) = \text{lub}(A)$
	$\|A\| := n \cdot \max_{i,j=1,\dots,n} \|a_{ij}\| = M(A)$
$\|x\|_2 := \sum_{i=1}^{n} \|x_i\|$	$\|A\|_2 := \max_{j=1,\dots,n} \sum_{i=1}^{n} \|a_{ij}\| = S(A) = \text{lub}(A)$
	$\|A\| := n \cdot \max_{i,j=1,\dots,n} \|a_{ij}\| = M(A)$
$\|x\|_3 := (\bar{x}^T x)^{1/2}$	$\|A\|_3 := (\Lambda_n)^{1/2} = H(A) = \text{lub}(A)$
	$\|A\| := n \max_{i,j=1,\dots,n} \|a_{ij}\| = M(A)$
	$\|A\| := \left(\sum_{i,j=1}^{n} \|a_{ij}\|^2\right)^{1/2} = N(A)$

Satz I.9 :

Es gelten die folgenden scharfen Abschätzungen:

1. $\|x\|_1 \leq \|x\|_3 \leq \|x\|_2$
2. $\|x\|_2 \leq \sqrt{n}\, \|x\|_3 \leq n\, \|x\|_1$

Für spätere Betrachtungen ist es wichtig, einige Aussagen über Folgen und Reihen von Matrizen zu machen.

Satz I.10 :

Für eine beliebige Matrix A gilt die Aussage :

$$A^n \xrightarrow[n\to\infty]{} 0 \quad \text{genau dann, wenn } \rho(A) < 1$$

Hinreichend für die Konvergenz gegen die Nullmatrix ist, daß für eine beliebige Matrixnorm gilt: $\|A\| < 1$

Satz I.11 :

Die geometrische Reihe $\sum_{k=0}^{n} A^k$ ist konvergent genau dann, wenn $A^n \xrightarrow[n\to\infty]{} 0$. Im Falle der Konvergenz gilt die folgende Beziehung :

$$\lim_{n\to\infty} \sum_{k=0}^{n} A^k = (E-A)^{-1}$$

Satz I.12 :

Gegeben seien eine multiplikative Matrixnorm und eine Matrix A mit $\|A\| < 1$, dann gilt die folgende Abschätzung :

$$\|(E \pm A)^{-1}\| \leq \frac{1}{1 - \|A\|}$$

§ 2. Grundlagen aus der Wahrscheinlichkeitstheorie

In diesem Paragraphen werden einige Sätze zusammengestellt, die in dieser Arbeit Anwendung finden.

Satz I.13 :

ξ_ν $\nu = 1,2,\ldots,n$ seien unabhängige zufällige Variable mit den Mittelwerten μ_ν und den Streuungen σ_ν . α_ν seien gegebene Zahlen.

$$\eta := \sum_{\nu=1}^{n} \alpha_\nu \xi_\nu$$

Dann gilt für den Mittelwert μ und die Streuung σ von η :

$$\mu = \sum_{\nu=1}^{n} \alpha_\nu \mu_\nu \; ; \quad \sigma^2 = \sum_{\nu=1}^{n} \alpha_\nu^2 \sigma_\nu^2$$

Sind insbesondere die zufälligen Variablen ξ_ν normalverteilt, so ist auch η normalverteilt.

Satz I.14 :

Die zufällige Variable ξ sei verteilt nach $N(0, \sigma^2)$.

$$\eta := |\xi|$$

Dann gilt für den Mittelwert $\bar{\mu}$ und die Streuung $\bar{\sigma}$ von η :

$$\bar{\mu} = \frac{2\sigma}{\sqrt{2\pi}} \; ; \quad \bar{\sigma} = \sigma \left(1 - \frac{2}{\pi}\right)^{1/2}$$

Beweis :

$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}}$ ist die Dichte von ξ . Für die Dichte $\varphi(x)$ von η gilt : $\varphi(x) = \begin{cases} 0 & \text{für } x \leq 0 \\ 2f(x) & \text{für } x > 0 \end{cases}$

Es gilt : $\bar{\mu} = \int_0^\infty x\varphi(x)dx = 2\int_0^\infty xf(x)dx = \frac{2}{\sigma\sqrt{2\pi}} \int_0^\infty xe^{-\frac{x^2}{2\sigma^2}} dx$

Durch Auswertung dieses Integrals erhält man den Wert für $\bar{\mu}$.

$$\bar{\sigma}^2 = \int_0^\infty (x-\bar{\mu})^2 \varphi(x)\,dx = 2\int_0^\infty (x-\bar{\mu})^2 f(x)\,dx =$$
$$= 2\Big[\int_0^\infty x^2 f(x)\,dx - 2\bar{\mu}\int_0^\infty x f(x)\,dx + \bar{\mu}^2 \int_0^\infty f(x)\,dx =$$
$$= 2\Big[\frac{\sigma^2}{2} - \bar{\mu}^2 + \frac{\bar{\mu}^2}{2}\Big] = \sigma^2\Big(1 - \frac{2}{\pi}\Big).$$

Satz I.15 :

$\xi_1, \xi_2, \ldots, \xi_n$ seien unabhängige nach $N(\mu_\nu, \sigma_\nu^2)$verteilte zufällige Variable mit den folgenden Eigenschaften :

1. $\mu_\nu > 0$ für $\nu = 1,2,\ldots,n$
2. Es existiert ein Index ν_0 mit $\mu_{\nu_0} = \max\limits_{\nu=1,\ldots,n} \mu_\nu$; $\sigma_{\nu_0} = \max\limits_{\nu=1,\ldots,n} \sigma_\nu$

Wir definieren die folgende zufällige Variable η :

$$\eta := \max_{\nu=1,\ldots,n} |\xi_\nu|$$

$1/2 \leq p < 1$ sei eine gegebene Wahrscheinlichkeit. Durch die folgenden Relationen werden die Größen λ und λ^* definiert :

$$W(\{\eta \leq \lambda\}) = p \; ; \; W(\{|\xi_{\nu_0}| \leq \lambda^*\}) = \sqrt[n]{p}$$

Dann gilt die Abschätzung : $\mu_{\nu_0} < \lambda \leq \lambda^*$

Beweis :

Wegen $1/2 \leq p$ folgt unmittelbar $\mu_{\nu_0} < \lambda$

Aus der Unabhängigkeit der zufälligen Variablen $\xi_1, \xi_2, \ldots, \xi_n$ folgt :

$$p = W(\{\eta \leq \lambda\}) = \prod_{\nu=1}^n W(\{|\xi_\nu| \leq \lambda\}) \leq \big[W(\{|\xi_{\nu_0}| \leq \lambda\})\big]^n$$

Daraus folgt : $W(\{|\xi_{\nu_0}| \leq \lambda\}) \leq \sqrt[n]{p}$

Für λ^* mit $W(\{|\xi_{\nu_0}| \leq \lambda^*\}) = \sqrt[n]{p}$ gilt dann $\lambda^* \geq \lambda$.

Zusatz :

$f(x)$ sei die Dichte der zufälligen Variablen ξ_{ν_0} .

Wir machen den Ansatz $\int_{\mu_{\nu_0}-c\,\sigma_{\nu_0}}^{\mu_{\nu_0}+c\,\sigma_{\nu_0}} f(x)\,dx = \sqrt[n]{P}$

Der Wert $c = c(\sqrt[n]{P})$ läßt sich unmittelbar einer Tabelle über Normalverteilung entnehmen. Wegen $\mu_{\nu_0} > 0$ gilt: $\mu_{\nu_0} + c\,\sigma_{\nu_0} \geq 2^*$
Für große n und $p \approx 1$ empfiehlt sich die bekannte Näherungsformel $\sqrt[n]{P} \approx 1 - \frac{1-p}{n}$.

Satz I.16 :

$x^T = (\xi_1, \xi_2, \ldots, \xi_n)$ sei ein zufälliger Vektor, dessen Komponenten unabhängig nach $N(0,\sigma^2)$ verteilt sind. $a_1 \neq \vartheta$ und $a_2 \neq \vartheta$ seien Vektoren.

$$\gamma := \frac{|a_1^T a_2|}{\|a_1\|_3 \|a_2\|_3}$$

Dann sind für $\gamma = 0$ die zufälligen Variablen $\xi_1 = a_1^T x$ und $\xi_2 = a_2^T x$ unabhängig.

Beweis : Vergleiche Schmetterer [5] S. 93.

Wir wollen Aussagen machen über den Grad der Abhängigkeit der zufälligen Variablen ξ_1 und ξ_2 für $0 < \gamma < 1$. Man wird erwarten, daß für $0 < \gamma << 1$ die zufälligen Variablen ξ_1 und ξ_2 "fast unabhängig" sind. Der Klärung dieser Frage gilt der folgende Satz.

Satz I.17 :

Wir wählen die Bezeichnungen von Satz I.16. Es sei $0 < \gamma < 1$. Dann existiert für die zufälligen Variablen ξ_1 und ξ_2 eine Darstellung $\xi_1 = \eta_1 + \xi_1 ; \xi_2 = \eta_2 + \xi_2$, für die folgendes gilt :

I. η_1 ist verteilt nach $N(0, \sigma_{11}^2)$
ξ_1 " " " $N(0, \sigma_{12}^2)$
η_2 " " " $N(0, \sigma_{21}^2)$
ξ_2 " " " $N(0, \sigma_{22}^2)$

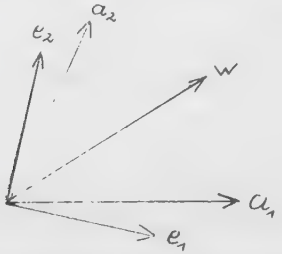
II. Es gelten die folgenden Abschätzungen :

$$\sigma \frac{\sqrt{2}}{2} \|a_1\|_3 < \sigma_{11} \; ; \; \sigma_{12} < \gamma \|a_1\|_3 \sigma$$

$$\sigma \frac{\sqrt{2}}{2} \|a_2\|_3 < \sigma_{21} \; ; \; \sigma_{22} < \gamma \|a_2\|_3 \sigma$$

Beweis :

Zum leichteren Verständnis des Beweises lasse man sich von der folgenden Skizze leiten :



o.B.d.A. $a_1^T a_2 > 0$

Es existieren zwei orthogonale normierte Vektoren e_1 und e_2 mit den folgenden Eigenschaften :

1. Es gilt die Darstellung :

$$\frac{a_1}{\|a_1\|_3} = \lambda_1 e_1 + \mu_2 e_2 \quad ; \quad \lambda_1^2 + \mu_2^2 = 1$$

$$\frac{a_2}{\|a_2\|_3} = \mu_1 e_1 + \lambda_2 e_2 \quad ; \quad \mu_1^2 + \lambda_2^2 = 1$$

2. $$\frac{\sqrt{\lambda}}{2} < |\lambda_1| = \frac{|a_1^T e_1|}{\|a_1\|_3}$$

Geometrisch bedeutet diese Ungleichung, daß a_1 in dem von e_1 und der Winkelhalbierenden w bestimmten Teilraum liegt.

$$\frac{\sqrt{\lambda}}{2} < |\lambda_2| = \frac{|a_2^T e_2|}{\|a_2\|_3}$$

Analog liegt a_2 in dem von e_2 und der Winkelhalbierenden w bestimmten Teilraum.

Für e_1 und e_2 gelten nach Konstruktion die Ungleichungen :

$$\mu_1 = \frac{|a_2^T e_1|}{\|a_2\|_3} < \frac{|a_1^T a_2|}{\|a_1\|_3 \|a_2\|_3} = \gamma \; ; \quad \mu_2 = \frac{|a_1^T e_2|}{\|a_1\|_3} < \frac{|a_1^T a_2|}{\|a_1\|_3 \|a_2\|_3} = \gamma .$$

Für die zufälligen Variablen ξ_1 und ξ_2 erhält man die Darstellung :

$$\xi_1 = \lambda_1 \|a_1\|_3 \cdot e_1^T x + \mu_2 \|a_1\|_3 \; e_2^T x$$

$$\xi_2 = \mu_1 \|a_2\|_3 \; e_1^T x + \lambda_2 \|a_2\|_3 \; e_2^T x$$

Nach Satz I.13 und Satz I.16 sind die zufälligen Variablen $e_1^T x$ und $e_2^T x$ unabhängige nach $N(0, \sigma^2)$ verteilte zufällige Variable.

Wir setzen : $\eta_1 := \lambda_1 \|a_1\|_3 \, e_1^T x \quad ; \quad \xi_1 := \mu_2 \|a_1\|_3 \, e_2^T x$

$\eta_2 := \lambda_2 \|a_2\|_3 \, e_2^T x \quad ; \quad \xi_2 := \mu_1 \|a_2\|_3 \, e_1^T x$

Berücksichtigt man die angegebenen Umbezeichnungen für $\lambda_1, \lambda_2, \mu_1, \mu_2$ so erhält man unmittelbar die Abschätzungen des Satzes.

Folgerung :

Für $\gamma << 1$ sind die zufälligen Variablen ξ_1, ξ_2 lediglich Störglieder und es gilt : $\xi_1 \approx \eta_1 \; ; \; \xi_2 \approx \eta_2$

d.h. ξ_1 und ξ_2 sind praktisch unabhängig.

II . K A P I T E L .

Untersuchungen über den Einfluss von Fehlern in den Ausgangsdaten eines linearen Gleichungssystems auf die Lösung.

§ 1. Abschätzungen für den maximalen Defektvektor bei Ungenauigkeiten in den Ausgangsdaten.

Wir gehen aus von dem numerisch gegebenen linearen Gleichungssystem

$$Ax = b \qquad \det A \neq 0$$

x sei die explizit vorliegende exakte Lösung des Gleichungssystems.

Eine Änderung der Ausgangsdaten bewirkt eine Änderung der Lösung.

$$\begin{aligned} A &\longrightarrow A+R \qquad \det(A+R) \neq 0 \\ b &\longrightarrow b+f \\ x &\longrightarrow x+\delta x \end{aligned}$$

Dabei gilt die folgende Beziehung :

$$(A+R)(x+\delta x) = b+f$$

Durch Auflösung dieser Gleichung nach δx erhält man hieraus :

$$\delta x = (A+R)^{-1}(f-Rx)$$

Die Matrizen R und die Vektoren f lassen wir innerhalb gegebener Mengen variieren.

$$M_n \supset \mathfrak{R} = \{ R = (r_{ij}) \} \qquad \det (A+R) \neq 0 \quad \text{für alle } R \in \mathfrak{R}$$

$$V_n \supset \mathfrak{F} = \{ f = (f_i) \}$$

M_n = Banachraum der (n,n)-Matrizen.

V_n = Banachraum der n-dimensionalen Vektoren.

z.B. : $\mathfrak{R} := \{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon \}$

$\mathfrak{F} := \{ f = (f_i) \mid |f_i| \leq \delta \}$

Wir interessieren uns für die Menge $\mathcal{G}$ der Gleichungssysteme :

$$\mathcal{G} := \{ g \equiv (A+R)(x+\delta x) = b+f \mid R \in \mathcal{R}, f \in \mathcal{F} \}$$

Im obigen Beispiel stellt die Menge $\mathcal{G}$ für hinreichend kleine Werte ε, δ eine Schar von linearen Gleichungssystemen dar, die zum gegebenen Gleichungssystem $Ax = b$ benachbart ist.

Wir fassen die innerhalb der gegebenen Fehlergrenzen auftretenden Abweichungen von der Lösung des linearen Gleichungssystems $Ax = b$ zu einer Menge zusammen.

$$\mathcal{V} := \{ \delta x \in V_n \mid x+\delta x \text{ ist Lösung von } g \in \mathcal{G} \}$$

Man beachte, dass wegen $\det(A+R) \neq 0$ die Menge $\mathcal{V}$ wohldefiniert ist. $\mathcal{V}$ nennen wir die Menge der Defektvektoren.

Wir interessieren uns nun für die folgende Größe :

$$\rho_i := \sup_{\delta x \in \mathcal{V}} \| \delta x \|_i$$

ρ_i gibt an, wie stark innerhalb der Fehlergrenzen, die durch $\mathcal{R}$ und $\mathcal{F}$ bestimmt sind, die Lösung eines "benachbarten" linearen Gleichungssystems bezüglich der Vektornorm $\| \ \|_i$ von der Lösung des gegebenen Gleichungssystems maximal abweichen kann.

Ziel der folgenden Ausführungen ist es, diese Größe ρ_i nach oben und - wenn möglich - nach unten abzuschätzen. Dabei werden wir uns auf die beiden folgenden Fehlertypen beschränken :

I. $\mathcal{R} := \{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon_{ij} \}$; $\mathcal{F} := \{ f = (f_i) \mid |f_i| \leq \delta_i \}$

II. $\mathcal{R} := \{ R \mid \|R\|_i \leq \varepsilon \}$; $\mathcal{F} := \{ f \mid \|f\|_i \leq \delta \}$

In beiden Fällen haben wir es mit kompakten Mengen $\mathcal{R}, \mathcal{F}$ zu tun. Daraus folgt insbesondere $\rho_i < \infty$.

Die Grundlage für Abschätzungen der Größe ρ_i nach oben bildet der folgende Satz :

Satz II.1 :

Gegeben seien das lineare Gleichungssystem $Ax = b$ mit $\det A \neq 0$

und die kompakten Mengen $\mathfrak{R}$, $\mathfrak{F}$. Es sei $\|A^{-1}\|_i \max_{R\in\mathfrak{R}} \|R\|_i < 1$.

Dann gelten für ρ_i die Abschätzungen (A1) , (A2) , (A3) , (A4) .

$$\rho_i \leq \frac{\max_{R,f} \|A^{-1}(f-Rx)\|_i}{1 - \max_{R\in\mathfrak{R}} \|A^{-1}R\|_i} = (A1)$$

$$\leq \frac{\max_{R,f} \|f-Rx\|_i}{\frac{1}{\|A^{-1}\|_i} - \max_{R\in\mathfrak{R}} \|R\|_i} = (A2)$$

$$\leq \frac{\max_{R\in\mathfrak{R}}\|Rx\|_i + \max_{f\in\mathfrak{F}}\|f\|_i}{\frac{1}{\|A^{-1}\|_i} - \max_{R\in\mathfrak{R}} \|R\|_i} = (A3)$$

$$\leq \frac{\max_{R\in\mathfrak{R}}\|R\|_i \, \|x\|_i + \max_{f\in\mathfrak{F}}\|f\|_i}{\frac{1}{\|A^{-1}\|_i} - \max_{R\in\mathfrak{R}}\|R\|_i} = (A4)$$

<u>Beweis</u> :

Für $R \in \mathfrak{R}$ und $f \in \mathfrak{F}$ gilt für den zugehörigen Defektvektor δx :

$$\delta x = (A+R)^{-1}(f-Rx)$$

Daraus folgt : $\|\delta x\|_i = \|(A+R)^{-1}(f-Rx)\|_i = \|(E+A^{-1}R)^{-1}A^{-1}(f-Rx)\|_i \leq$

$$\leq \|(E+A^{-1}R)^{-1}\|_i \, \|A^{-1}(f-Rx)\|_i \ .$$

Wegen $\|A^{-1}R\|_i \leq \|A^{-1}\|_i \|R\|_i \leq \|A^{-1}\|_i \max_{R\in\mathfrak{R}}\|R\|_i < 1$ gilt nach Satz I.12 :

$$\|E+A^{-1}R\|_i \leq \frac{1}{1 - \|A^{-1}R\|_i}$$

Daraus folgt : $\|\delta x\|_i \leq \frac{\|A^{-1}(f-Rx)\|_i}{1 - \|A^{-1}R\|_i}$

Daraus folgt : $\rho_i = \max_{\delta x\in\mathfrak{d}} \|\delta x\|_i \leq \frac{\max_{R,f}\|A^{-1}(f-Rx)\|_i}{1 - \max_{R\in\mathfrak{R}}\|A^{-1}R\|_i}$

Die weiteren Abschätzungen folgen unmittelbar unter Ausnützung elementarer Eigenschaften von Vektor- und Matrixnormen.

Wir haben das Ziel, Abschätzungen für ρ_i zu finden. Die Methode hierfür besteht darin, die Abschätzungen des letzten Satzes für spezielle Mengen $\mathcal{R}, \mathcal{F}$ auszuwerten. Es ist notwendig, zunächst eine Reihe vorbereitender Hilfssätze zu beweisen.

Hilfssatz II.1 :

a,b seien zwei beliebige Vektoren, $\|\ \|$ eine beliebige Vektornorm. Dann gilt mindestens eine der beiden folgenden Relationen:

$$\|a+b\| \geq \|a\| \ , \ \|a-b\| \geq \|a\|$$

Beweis :

Annahme : $\|a+b\| < \|a\|$ und $\|a-b\| < \|a\|$. Dann gilt:

$$2\|a\| = \|a+b+a-b\| \leq \|a+b\| + \|a-b\| < 2\|a\|$$

Dies ist ein Widerspruch. Daraus folgt die Behauptung.

Hilfssatz II.2 :

a,b seien zwei Vektoren mit $\|a+b\| \geq \|a\|$. Dann gilt:

$$\|a+\lambda b\| \geq \|a+b\| \text{ für } \lambda \geq 1$$

Beweis :

Für $\lambda \geq 1$ gilt : $\lambda\|a+b\| = \|\lambda a+\lambda b\| = \|(\lambda-1)a+a+\lambda b\| \leq$

$$\leq (\lambda-1)\|a\| + \|a+\lambda b\| \leq (\lambda-1)\|a+b\| + \|a+\lambda b\|$$

Daraus folgt : $\|a+\lambda b\| \geq \|a+b\|$

Der folgende Hilfssatz ist ein Spezialfall der bekannten Aussage, daß ein konvexes Funktional auf einem Polyeder sein Maximum in einer Ecke annimmt.

Hilfssatz II.3 :

$A = (a_1\ a_2\ \dots\ a_n)$ sei eine beliebige Matrix und $\|\ \|$ eine beliebige Vektornorm. $\tau_i \geq 0$ $i = 1,2,\dots,n$ seien gegebene Zahlen.

$$\mathcal{T}_1 := \left\{ t = (t_i) \,\middle|\, |t_i| \leq \tau_i \right\}$$

$$\mathcal{T}_2 := \left\{ t = (t_i) \,\middle|\, |t_i| = \tau_i \right\}$$

Dann gilt : $\max\limits_{t \in \mathcal{T}_1} \|At\| = \max\limits_{t \in \mathcal{T}_2} \|At\|$; $\max\limits_{t \in \mathcal{T}_1} \|t\| = \max\limits_{t \in \mathcal{T}_2} \|t\|$

Beweis :

Da $\mathcal{T}_1$ kompakt ist, existiert ein Vektor $t_o = (t_i^{(o)}) \in \mathcal{T}_1$ mit

$$\|At_o\| = \max_{t \in \mathcal{T}_1} \|At\|$$

Die Beweisidee besteht darin, einen Vektor $t^* \in \mathcal{T}_2$ zu konstruieren, für den gilt : $\|At^*\| = \|At_o\|$

I. Gilt $t_o \in \mathcal{T}_2$, dann setze $t^* := t_o$

II. Gilt $t_o \notin \mathcal{T}_2$, dann existiert i_1 mit $|t_{i_1}| < \tau_{i_1}$

Man setze : $a := \sum_{i \neq i_1} t_i^{(o)} a_i$, $b := t_{i_1}^{(o)} a_{i_1}$

Es gilt : $\|a\| \leq \|a+b\| = \|At_o\|$

1. $b = \vartheta$

Nach Hilfssatz II.1 gilt mindestens eine der beiden Relationen :

$$\|a + \tau_{i_1} a_{i_1}\| \geq \|a\| \quad , \quad \|a - \tau_{i_1} a_{i_1}\| \geq \|a\| \quad ; \quad a = At_o$$

Setze $t_1 = (t_i^{(1)})$ $\begin{cases} t_i^{(1)} = t_i^{(o)} \text{ für } i \neq i_1 \\ t_{i_1} = \begin{cases} \tau_{i_1} \text{ , falls } \|a + \tau_{i_1} a_{i_1}\| \geq \|a\| \\ -\tau_{i_1} \text{ sonst} \end{cases} \end{cases}$

Nach Konstruktion gilt : $t_1 \in \mathcal{T}_1$ und $\|At_1\| = \|At_o\|$

2. $b \neq \vartheta$

Setze $\alpha := \dfrac{\tau_{i_1}}{|t_{i_1}^{(o)}|} > 1$

Setze $t_1 = (t_i^{(1)})$ mit $t_i^{(1)} = \begin{cases} t_i^{(o)} & \text{für } i \neq i_1 \\ \tau_{i_1} \operatorname{sgn} t_{i_1}^{(o)} & \text{für } i = i_1 \end{cases}$

Nach Konstruktion gilt : $t_1 \in \mathcal{T}_1$

Nach Hilfssatz II.2 gilt : $\|At_1\| = \|a + \alpha b\| \geq \|a+b\| = \|At_o\|$

(i) $t_1 \in \mathcal{T}_2$, so setze $t^* := t_1$

(ii) $t_1 \notin \mathcal{T}_2$, so führe man die eben angegebene Konstruktion mit einem Index i_2 durch. Nach endlich vielen Schritten

gelangt man zu einem Vektor $t^* \in \mathfrak{Y}_2$ mit $\|At^*\| = \|At_o\|$.

Für A = E erhält man unmittelbar die zweite Aussage des Hilfssatzes.

Bemerkung :

Anschaulich besagt Hilfssatz II.3 , daß das Funktional $\|At\|$ in einem Quader sein Maximum in einer Ecke annimmt.

Hilfssatz II.4 :

Wir wählen die Voraussetzungen von Hilfssatz II.3 . Außerdem sei

$A^* := (\tau_1 a_1 \ \tau_2 a_2 \cdots \tau_n a_n)$ und $\mathfrak{E} := \{ e = (e_i) \mid |e_i| = 1 \}$

Dann gilt : $$\max_{t \in \mathfrak{Y}_1} \|At\| = \max_{e \in \mathfrak{E}} \|A^* e\| \leq \max_{i=1,\dots,n} \tau_i \max_{e \in \mathfrak{E}} \|Ae\|$$

$$\max_{t \in \mathfrak{Y}_1} \|t\| \leq \max_{i=1,\dots,n} \tau_i \max_{e \in \mathfrak{E}} \|e\|$$

Gilt speziell $\tau = \tau_1 = \tau_2 = \cdots = \tau_n$, so erhält man :

$$\max_{t \in \mathfrak{Y}_1} \|At\| = \tau \max_{e \in \mathfrak{E}} \|Ae\|$$

$$\max_{t \in \mathfrak{Y}_1} \|t\| = \tau \max_{e \in \mathfrak{E}} \|e\|$$

Beweis :

Die Aussage ergibt sich als Folgerung von Hilfssatz II.3 .

Die Aussagen des folgenden Hilfssatzes ermöglichen eine kürzere und übersichtlichere Beweisführung bei späteren Sätzen.

Hilfssatz II.5 :

I. a = (a_i) sei ein gegebener Vektor ; $\mathfrak{Y} := \{ t = (t_i) \mid |t_i| \leq \tau_i \}$

Dann gilt : $$\{ a^T t \mid t \in \mathfrak{Y} \} = \{ \alpha \mid |\alpha| \leq \sum_{i=1}^{n} \tau_i |a_i| \}$$

Gilt speziell $\tau = \tau_1 = \tau_2 = \cdots = \tau_n$, so erhält man :

$$\{ a^T t \mid t \in \mathfrak{Y} \} = \{ \alpha \mid |\alpha| \leq \tau \|a\|_2 \}$$

II. a = (a_i) sei ein gegebener Vektor ; $\mathfrak{M} := \{ M = (m_{ij}) \mid |m_{ij}| \leq \tau_{ij} \}$

Dann gilt : $$\{ Ma \mid M \in \mathfrak{M} \} = \{ t = (t_i) \mid |t_i| \leq \sum_{j=1}^{n} \tau_{ij} |a_j| \}$$

Gilt : $\tau_{ij} = \tau \geq 0 \quad i,j = 1,\ldots,n$, so erhält man :

$$\{ Ma \mid M \in \mathfrak{M} \} = \{ t = (t_i) \mid |t_i| \leq \tau \|a\|_2 \}$$

III. $\gamma := \{ s = (s_i) \mid \|s\| = 1 \}$; $\tau := \{ t = (t_i) \mid |t_i| \leq \tau_i \}$

Dann gilt : $\{ t^T s \mid t \in \tau , s \in \gamma \} = \{ \alpha \mid |\alpha| \leq \max\limits_{\|s\|=1} \sum\limits_{i=1}^{n} \tau_i |s_i| \}$

Gilt : $\tau = \tau_1 = \tau_2 = \cdots = \tau_n$, so erhält man :

$$\{ t^T s \mid t \in \tau , s \in \gamma \} = \{ \alpha \mid |\alpha| \leq \tau \max_{\|s\|=1} \|s\|_2 \}$$

IV. $\gamma := \{ s = (s_i) \mid \|s\| = 1 \}$; $\mathfrak{M} := \{ M = (m_{ij}) \mid |m_{ij}| \leq \tau_{ij} \}$

Dann gilt : $\{ Ms \mid M \in \mathfrak{M}, s \in \gamma \} \subset \{ t = (t_i) \mid |t_i| \leq \max\limits_{\|s\|=1} \sum\limits_{j=1}^{n} \tau_{ij} |s_j| \}$

Gilt : $\tau_{ij} = \tau \geq 0$ für $i,j = 1,\ldots,n$, so erhält man :

$$\{ Ms \mid M \in \mathfrak{M}, s \in \gamma \} = \{ t = (t_i) \mid |t_i| \leq \tau \max \|s\|_2 \}$$

Beweis :

Zu I.

Es gilt : $|a^T t| = \left| \sum\limits_{i=1}^{n} a_i t_i \right| \leq \sum\limits_{i=1}^{n} |a_i| |t_i| \leq \sum\limits_{i=1}^{n} \tau_i |a_i|$

Für $t = (t_i)$ mit $t_i = \operatorname{sgn} a_i \cdot \tau_i$ gilt : $a^T t = \sum\limits_{i=1}^{n} \tau_i |a_i|$

Daraus folgt unmittelbar die Gleichheit beider Mengen.

Zu II.

Für $t = (t_i) = Ma$ gilt : $|t_i| = \left| \sum\limits_{j=1}^{n} m_{ij} a_j \right| \leq \sum\limits_{j=1}^{n} \tau_{ij} |a_j|$

Für $M = (m_{ij})$ mit $m_{ij} = \operatorname{sgn} a_j \cdot \tau_{ij}$ gilt : $t_i = \sum\limits_{j=1}^{n} \tau_{ij} |a_j|$

Daraus folgt unmittelbar die Gleichheit beider Mengen.

Zu III.

Es gilt : $|t^T s| \leq \sum\limits_{i=1}^{n} \tau_i |s_i| \leq \max\limits_{\|s\|=1} \sum\limits_{i=1}^{n} \tau_i |s_i| = \sum\limits_{i=1}^{n} \tau_i |s_i^{(o)}|$

Für $s_o = (s_i^{(o)})$ und $t = (t_i)$ mit $t_i = \operatorname{sgn} s_i^{(o)} \tau_i$ gilt :

$$|t^T s_o| = \max_{\|s\|=1} \sum_{i=1}^{n} \tau_i |s_i|$$

Daraus folgt unmittelbar die Gleichheit beider Mengen.

Zu IV.

Für $t = (t_i) = Ms$ gilt : $|t_i| \leq \left| \sum\limits_{j=1}^{n} m_{ij} s_j \right| \leq \max\limits_{\|s\|=1} \sum\limits_{j=1}^{n} \tau_{ij} |s_j|$

Für verschiedene Indices i wird das Maximum im allgemeinen für verschiedene Vektoren s angenommen. Daraus folgt unmittelbar die angegebene Inklusion. Die Gleichheit der Mengen für $\tau = \tau_{ij}$ ist klar.

Wir haben jetzt die Voraussetzungen geschaffen, die Größe ρ_i für einige spezielle Fehlermengen $\mathcal{R}$ und $\mathcal{F}$ abzuschätzen.

Satz II.2 :

Gegeben seien das Gleichungssystem $Ax = b$ mit $\det A \neq 0$ und die Mengen $\mathcal{R} := \{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon \}$; $\mathcal{F} := \{ f = (f_i) \mid |f_i| \leq \delta \}$

$$\alpha_i := \frac{1}{\max\limits_{e \in \mathcal{E}} \|A^{-1} e\|_i} \quad ; \quad \alpha_i - \varepsilon \max_{\|s\|_i = 1} \|s\|_2 > 0$$

Dann gelten die folgenden Abschätzungen :

$$\underline{\varphi}_i(A, b, \varepsilon, \delta) := \frac{\varepsilon \|x\|_2 + \delta}{\alpha_i + \varepsilon \max\limits_{\|s\|_i = 1} \|s\|_2} \leq \rho_i \leq \frac{\varepsilon \|x\|_2 + \delta}{\alpha_i - \varepsilon \max\limits_{\|s\|_i = 1} \|s\|_2} =: \bar{\varphi}_i(A, b, \varepsilon, \delta)$$

Beweis :

I. Nachweis der Abschätzung $\rho_i \leq \bar{\varphi}_i$:

Wir gehen aus von Abschätzung (A1) von Satz II.1 .

1. Auswertung von $\max\limits_{R, f} \|A^{-1}(f - Rx)\|_i$:

Nach Hilfssatz II.5 erhält man die folgende Beziehung :

$$\{ f - Rx \mid f \in \mathcal{F}, R \in \mathcal{R} \} = \{ t = (t_i) \mid |t_i| \leq \varepsilon \|x\|_2 + \delta \}$$

Nach Hilfssatz II.4 folgt hieraus unmittelbar :

$$\max_{R, f} \|A^{-1}(f - Rx)\|_i = \left[\varepsilon \|x\|_2 + \delta \right] \max_{e \in \mathcal{E}} \|A^{-1} e\|_i$$

2. Auswertung von $\max\limits_{R \in \mathcal{R}} \|A^{-1} R\|_i = \max\limits_{R, \|s\|_i = 1} \|A^{-1} Rs\|_i$:

Nach Hilfssatz II.5 erhält man die folgende Beziehung :

$$\{ Rs \mid R \in \mathcal{R}, \|s\|_i = 1 \} = \{ t = (t_i) \mid |t_i| \leq \varepsilon \max_{\|s\|_i = 1} \|s\|_2 \}$$

Nach Hilfssatz II.4 folgt hieraus unmittelbar :

$$\max_{R, \|s\|_i = 1} \|A^{-1} Rs\|_i = \varepsilon \max_{\|s\|_i = 1} \|s\|_2 \max_{e \in \mathcal{E}} \|A^{-1} e\|_i$$

Setzt man diese Terme in die Abschätzung (A1) ein, so erhält man nach elementarer Umformung die Abschätzung $\bar{\varphi}_i$.

II. Nachweis der Abschätzung $\underline{\varphi}_i \leq \rho_i$

Beweisidee : Wir konstruieren einen Vektor $\delta x \in \mathcal{V}$, für den gilt : $\underline{\varphi}_i \leq \|\delta x\|_i \leq \rho_i$

$\delta x \in \mathcal{V} \iff (A+R)(x+\delta x) = b+f \quad , \quad R \in \mathcal{R} \; , \; f \in \mathcal{F} \iff$

$A\delta x + R\delta x + Rx - f = \mathcal{V}, \; R \in \mathcal{R} \; , \; f \in \mathcal{F}$

Ansatz: $\delta x = \lambda s \quad \|s\|_i = 1$

Wir wählen die Größen s,R,f folgendermaßen:

$$s := \frac{A^{-1}e_o}{\|A^{-1}e_o\|_i} \quad , \text{ wobei gilt : } \alpha_i = \frac{1}{\|A^{-1}e_o\|_i} \qquad e_o = (e_i^{(o)})$$

$R = (r_{kl})$ mit $r_{kl} = -\varepsilon e_k^{(o)} \operatorname{sgn} x_l$

$f = (f_k)$ mit $f_k = \delta e_k^{(o)}$

Man beachte : $\|s\|_i = 1 \quad ; \quad R \in \mathcal{R} \; , \; f \in \mathcal{F}$

Für die j-te Komponente des Vektors $\lambda As + \lambda Rs + Rx - f$ erhält man:

$$e_j^{(o)} \left(\lambda \alpha_i - \varepsilon \lambda \sum_{\nu=1}^{n} s_\nu \operatorname{sgn} x_\nu - \varepsilon \|x\|_2 - \delta \right)$$

Da der Klammerausdruck nicht von j abhängt, läßt sich ein $\lambda > 0$ angeben, so daß alle Komponenten des obigen Vektors verschwinden.

Man wähle : $$\lambda := \frac{\varepsilon \|x\|_2 + \delta}{\alpha_i - \varepsilon \sum_{\nu=1}^{n} s_\nu \operatorname{sgn} x_\nu}$$

Nach Konstruktion gilt die Gleichung : $\lambda As + \lambda Rs + Rx - f = \mathcal{V}$

Daraus folgt : $\lambda s \in \mathcal{V}$. Ferner gilt : $\|\lambda s\|_i = \lambda \geq \underline{\varphi}_i$

Hieraus folgt die angegebene Abschätzung für ϱ_i nach unten.

Folgerung :

Für $\varepsilon \max_{\|s\|_i = 1} \|s\|_2 << \alpha_i$ sind die Abschätzungen $\underline{\varphi}_i$ und $\bar{\varphi}_i$ sehr gut.

Zusatz :

Für die Normen $i = 1,2,3$ läßt sich der Ausdruck $\max_{\|s\|_i = 1} \|s\|_2$ unmittelbar auswerten. Es gilt :

$$\max_{\|s\|_1 = 1} \|s\|_2 = n \quad ; \quad \max_{\|s\|_2 = 1} \|s\|_2 = 1 \quad ; \quad \max_{\|s\|_3 = 1} \|s\|_2 = \sqrt{n}$$

Daraus ergeben sich für die Normen $i = 1,2,3$ die Abschätzungen:

$$\underline{\varphi}_1(A,b,\varepsilon,\delta) := \frac{\varepsilon \|x\|_2 + \delta}{\alpha_1 + \varepsilon \sqrt{n}} \leq \varrho_1 \leq \frac{\varepsilon \|x\|_2 + \delta}{\alpha_1 - \varepsilon n} =: \bar{\varphi}_1(A,b,\varepsilon,\delta)$$

$$\underline{\varphi}_2(A,b,\varepsilon,\delta) := \frac{\varepsilon \|x\|_2 + \delta}{\alpha_2 + \varepsilon} \leq \varrho_2 \leq \frac{\varepsilon \|x\|_2 + \delta}{\alpha_2 - \varepsilon} =: \bar{\varphi}_2(A,b,\varepsilon,\delta)$$

$$\underline{\varphi}_3(A,b,\varepsilon,\delta) := \frac{\varepsilon\|x\|_2 + \delta}{\alpha_3 + \varepsilon\sqrt{n}} \leq \rho_3 \leq \frac{\varepsilon\|x\|_2 + \delta}{\alpha_3 - \varepsilon\sqrt{n}} =: \bar{\varphi}_3(A,b,\varepsilon,\delta)$$

Der folgende Satz gibt Auskunft über das Verhalten der Größen ρ_i, $\bar{\varphi}_i$, $\underline{\varphi}_i$ für $\varepsilon,\delta \longrightarrow 0$

Satz II.3 :

Wir gehen aus von den Voraussetzungen und Bezeichnungen von Satz II.2. Es sei $\delta = c\cdot\varepsilon$. Dann gilt für $\varepsilon \to 0$:

$$\rho_i = \underline{\varphi}_i + O(\varepsilon^2) \; ; \; \rho_i = \bar{\varphi}_i + O(\varepsilon^2)$$

Beweis :

$$\frac{|\bar{\varphi}_i - \underline{\varphi}_i|}{\varepsilon^2} = \frac{2 \max\limits_{\|s\|_i=1}\|s\|_2 \left[\|x\|_2 + c\right]}{\alpha^2 - \left[\varepsilon \max\limits_{\|s\|_i=1}\|s\|_2\right]^2} \xrightarrow{\varepsilon \to 0} \text{const.}$$

Hieraus folgt unmittelbar die Behauptung des Satzes.

Die Ergebnisse von Satz II.2 wollen wir durch ein einfaches numerisches Beispiel veranschaulichen.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 5 & 7 & 8 \\ 7 & 8 & 11 \end{pmatrix} \; ; \; b = \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} \; ; \quad \varepsilon = 0{,}1 \; ; \quad \delta = 0{,}1$$

Wir werden nun für dieses Zahlenbeispiel bezüglich der Norm $i = 3$ die Abschätzungen $\underline{\varphi}_3$ und $\bar{\varphi}_3$ berechnen.

$$A^{-1} = \begin{pmatrix} -1{,}08 & -0{,}17 & 0{,}42 \\ -0{,}08 & 0{,}83 & -0{,}58 \\ 0{,}75 & -0{,}50 & 0{,}25 \end{pmatrix} \; ; \quad x = \begin{pmatrix} -7{,}67 \\ 1{,}33 \\ 0{,}50 \end{pmatrix}$$

s_1, s_2, s_3 seien die Spaltenvektoren von A^{-1}. Die Berechnung der Vektoren $A^{-1}e$ $e \in \mathcal{E}$ ergibt :

$$d_1 = s_1 + s_2 + s_3 = \begin{pmatrix} -0{,}73 \\ 0{,}17 \\ 0{,}50 \end{pmatrix} \text{ Dann gilt : } \|d_1\|_3 = 1{,}03$$

$$d_2 = s_1 + s_2 - s_3 = \begin{pmatrix} -1{,}67 \\ 1{,}33 \\ 0{,}00 \end{pmatrix} \text{ Dann gilt : } \|d_2\|_3 = 2{,}14$$

$$d_3 = s_1 - s_2 + s_3 = \begin{pmatrix} -0,49 \\ -1,49 \\ 1,50 \end{pmatrix} \qquad \text{Dann gilt : } \|d_3\|_3 = 2,17$$

$$d_4 = s_1 - s_2 - s_3 = \begin{pmatrix} -1,33 \\ -0,33 \\ 1,00 \end{pmatrix} \qquad \text{Dann gilt : } \|d_4\|_3 = 1,97$$

Daraus ergibt sich : $= \frac{1}{\max \|A^{-1} e\|_3} = \frac{1}{2,17} = 0,47$

$\underline{\rho}_3 = \frac{\varepsilon \|x\|_2 + \delta}{\alpha_3 + \varepsilon \sqrt{n}} = 1,64$; $\overline{\rho}_3 = \frac{\varepsilon \|x\|_2 + \delta}{\alpha_3 - \varepsilon \sqrt{n}} = 3,50$

Daraus erhält man die Abschätzung : $1,64 \leq \rho_3 \leq 3,50$

Bemerkung :

Wir haben hier eine Matrix mit schlechter Kondition vorliegen. Trotz kleiner Fehlerschranken ε , δ können recht große Abweichungen in der Lösung auftreten.

Die Auswertung der Abschätzung (A2) von Satz II.1 liefert uns eine weitere Abschätzung für die Größe ρ_i.

Satz II.4 :

Gegeben seien das Gleichungssystem $Ax = b$ $\det A \neq 0$,

$$\mathcal{R} := \left\{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon \right\}, \quad \mathcal{F} := \left\{ f = (f_i) \mid |f_i| \leq \delta \right\}$$

$$\beta_i := \frac{1}{\|A^{-1}\|_i \max_{\|e\|_2 \leq \sqrt{n}} \|e\|_i} \quad ; \quad \beta_i - \varepsilon \max_{\|s\|_2 \leq \sqrt{n}} \|s\|_2 > 0$$

Dann gilt die folgende Abschätzung :

$$\rho_i \leq \frac{\varepsilon \|x\|_2 + \delta}{\beta_i - \varepsilon \max \|s\|_2} =: F_i(A, b, \varepsilon, \delta)$$

Beweis :

Wir gehen aus von Abschätzung (A2) von Satz II.1 .

1. Auswertung von $\max_{R,f} \|f - Rx\|_i$:

Nach Hilfssatz $II.5_{II}$ erhält man die folgende Beziehung :

$$\left\{ f - Rx \mid f \in \mathcal{F} , R \in \mathcal{R} \right\} \quad \left\{ t = (t_i) \mid |t_i| \leq \varepsilon \|x\|_2 + \delta \right\}$$

Nach Hilfssatz II.4 folgt hieraus unmittelbar :

$$\max_{R,f} \|f-Rx\|_i = \left[\varepsilon \|x\|_2 + \delta \right] \max_{e \in \mathcal{E}} \|e\|_i$$

2. Auswertung von $\max_{R \in \mathcal{R}} \|R\|_i = \max_{R, \|Rs\|_i = 1} \|Rs\|_i$

Nach Hilfssatz II.5 erhält man die folgende Beziehung :

$$\left\{ Rs \;\middle|\; R \in \mathcal{R} \;,\; \|s\|_i = 1 \right\} = \left\{ t = (t_i) \;\middle|\; |t_i| \leq \varepsilon \max_{\|s\|_i = 1} \|s\|_2 \right\}$$

Nach Hilfssatz II.4 folgt hieraus unmittelbar :

$$\max_{R, \|Rs\|_i = 1} \|Rs\|_i = \varepsilon \max_{\|s\|_i = 1} \|s\|_2 \max_{e \in \mathcal{E}} \|e\|_i$$

Setzt man diese Terme in die Abschätzung (A2) ein, so erhält man nach elementarer Umformung die Abschätzung $\mathcal{F}_i$.

Zusatz :

Für die Normen $i = 1,2,3$ läßt sich der Ausdruck $\max_{e \in \mathcal{E}} \|e\|_i$ unmittelbar auswerten. Es gilt :

$$\max_{e \in \mathcal{E}} \|e\|_1 = 1 \quad ; \quad \max_{e \in \mathcal{E}} \|e\|_2 = n \quad ; \quad \max_{e \in \mathcal{E}} \|e\|_3 = \sqrt{n}$$

Daraus ergeben sich für die Normen $i = 1,2,3$ die Abschätzungen :

$$\rho_1 \leq \frac{\varepsilon \|x\|_2 + \delta}{\frac{1}{\|A^{-1}\|_1} - \varepsilon n} =: F_1(A,b,\varepsilon,\delta)$$

$$\rho_2 \leq \frac{\varepsilon \|x\|_2 + \delta}{\frac{1}{n\|A^{-1}\|_2} - \varepsilon} =: F_2(A,b,\varepsilon,\delta)$$

$$\rho_3 \leq \frac{\varepsilon \|x\|_2 + \delta}{\frac{1}{\sqrt{n}\|A^{-1}\|_3} - \varepsilon\sqrt{n}} =: F_3(A,b,\varepsilon,\delta)$$

Bemerkung :

Die Abschätzung F_i ist schlechter als φ_i ; allerdings - wie später genauer ausgeführt wird - einer numerischen Auswertung im allgemeinen leichter zugänglich. Für die Normen $i = 1,2,3$ werden wir die Größen α_i, β_i und damit die Abschätzungen φ_i, $\mathcal{F}_i$ miteinander vergleichen.

Satz II.5 :

A mit $\det A \neq 0$ sei eine beliebige Matrix .

$$\alpha_i := \frac{1}{\max\limits_{e \in \mathfrak{E}} \|A^{-1}e\|_i} \quad ; \quad \beta_i := \frac{1}{\|A^{-1}\|_i \max\limits_{e \in \mathfrak{E}} \|e\|_i}$$

Dann gelten für die Normen $i = 1,2,3$ die folgenden Abschätzungen :

$$1 = \frac{\alpha_1}{\beta_1} \; ; \; 1 \leq \frac{\alpha_2}{\beta_2} \leq n \; ; \; 1 \leq \frac{\alpha_3}{\beta_3} < \sqrt{n}$$

Beweis :

I. $i = 1$

$$\frac{1}{\alpha_1} = \max_{e \in \mathfrak{E}} \|A^{-1}e\|_1 = \max_{\|s\|_1 = 1} \|A^{-1}s\|_1 = \|A^{-1}\|_1 = \frac{1}{\beta_1}$$

II. $i = 2$

$s_1, s_2, \ldots, s_n$ seien die Spaltenvektoren von A^{-1}. Dann existiert ein Index $1 \leq i_o \leq n$ mit $\|A^{-1}\|_2 = \|s_{i_o}\|_2$. Man sieht leicht die folgenden Relationen ein :

$$\|A^{-1}\|_2 = \|s_{i_o}\|_2 \leq \max_{e \in \mathfrak{E}} \|A^{-1}e\|_2 \leq n\|s_{i_o}\|_2 = n\,\|A^{-1}\|_2$$

Die beiden folgenden Beispiele zeigen, daß in den „$\leq$"- Zeichen die „$=$"- Zeichen eintreten können.

1. Beispiel : $A^{-1} = \begin{pmatrix} 5 & -2 \\ 3 & 2 \end{pmatrix}$. In diesem Beispiel gilt :

$$\|A^{-1}\|_2 = 8 \; ; \; \max_{e \in \mathfrak{E}} \|A^{-1}e\|_2 = 8$$

2. Beispiel : $A^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$. In diesem Beispiel gilt :

$$\|A^{-1}\|_2 = 4 \; ; \; \max_{e \in \mathfrak{E}} \|A^{-1}e\|_2 = 8 \; ; \; n = 2$$

Berücksichtigt man $\alpha_2 = \frac{1}{\max\limits_{e \in \mathfrak{E}} \|A^{-1}e\|_2}$ und $\beta_2 = \frac{1}{n\|A^{-1}\|_2}$,

so ergibt sich die im Satz angegebene Abschätzung.

III. $i = 3$

Man sieht leicht die folgenden Relationen ein :

$$\left\{ t = (t_i) \;\middle|\; |t_i| \leq \frac{1}{\sqrt{n}} \right\} \subset \left\{ t = (t_i) \;\middle|\; \|t\|_3 \leq 1 \right\} \subset \left\{ t = (t_i) \;\middle|\; |t_i| \leq 1 \right\}$$

Beachtet man Hilfssatz II.3, so erhält man :

1. $\|A^{-1}\|_3 = \max\limits_{\|s\|_3 = 1} \|A^{-1}s\|_3 < \max\limits_{e \in \mathfrak{E}} \|A^{-1}e\|_3$

2. $\max_{e\in\mathcal{E}} \|A^{-1}\frac{e}{\sqrt{n}}\|_3 \leq \max_{\|s\|_3=1} \|A^{-1}s\|_3 = \|A^{-1}\|_3$. Daraus folgt :

$$\max_{e\in\mathcal{E}} \|A^{-1}e\|_3 \leq \sqrt{n}\, \max_{\|s\|_3=1} \|A^{-1}s\|_3 = \sqrt{n}\, \|A^{-1}\|_3$$

Wir zeigen noch, daß beide Ungleichungen scharf sind.

Zu 1 : Es sei $B = (b_1\ b_2\ \dots\ b_n)$ mit $b_1 \neq \mathcal{O}$, $b_2 = \dots = b_n = \mathcal{O}$

Dann folgt : $\|B\|_3 = \max_{\|s\|_3=1} \|Bs\|_3 = \|b_1\|_3 = \max_{e\in\mathcal{E}} \|Be\|_3$

Je mehr also eine Spalte von A^{-1} gegenüber den anderen Spalten überwiegt, desto genauer wird die Ungleichung.

Zu 2 : Für $A^{-1} = E$ wird das „=“-Zeichen angenommen.

Wir gehen jetzt von den folgenden Fehlermatrizen aus

$$\mathcal{R} := \left\{ R = (r_{ij}) \;\middle|\; |r_{ij}| \leq \varepsilon\, l_{ij} (\geq 0) \right\}$$

$$\mathcal{F} := \left\{ f = (f_i) \;\middle|\; |f_i| \leq \varepsilon\, l_i (\geq 0) \right\}$$

und geben Abschätzungen für die Größe ρ_i an.

Satz II.6 :

Gegeben seien das lineare Gleichungssystem $Ax = b$ $\det A \neq 0$ und

$$\mathcal{R} := \left\{ R = (r_{ij}) \;\; |r_{ij}| \leq \varepsilon\, l_{ij} \right\} ; \mathcal{F} := \left\{ f = (f_i) \;\; |f_i| \leq \varepsilon\, l_i \right\}$$

$$\tau_\nu := l_\nu + \sum_{j=1}^{n} l_{\nu j}|x_j| > 0 \quad ; \quad \sigma_\nu^{(i)} := \max_{\|t\|_i=1} \sum_{j=1}^{n} l_{\nu j}|t_j|$$

$$A^{-1} := (s_1 \dots s_n) \; ; \; A^* := (\sigma_1^{(i)} s_1 \dots \sigma_n^{(i)} s_n) \; ; \; \bar{A} := (\tau_1 s_1 \dots \tau_n s_n)$$

$$\alpha_i^* := \frac{1}{\max_{e\in\mathcal{E}} \|A^* e\|_i} \quad ; \quad \bar{\alpha}_i := \frac{1}{\max_{e\in\mathcal{E}} \|\bar{A} e\|_i}$$

$$\bar{\alpha}_i - \varepsilon \frac{\bar{\alpha}_i}{\alpha_i^*} > 0 \; ; \; \bar{\alpha}_i - \varepsilon \max_{\nu=1,\dots,n} \frac{\sigma_\nu^{(i)}}{\tau_\nu} > 0$$

Dann gelten die folgenden Abschätzungen:

$$\rho_i \leq \frac{\varepsilon}{\bar{\alpha}_i - \varepsilon \frac{\bar{\alpha}_i}{\alpha_i^*}} =: S_i(A, b, \varepsilon, l_{ij}, l_i)$$

$$\rho_i \le \frac{\varepsilon}{\bar{\alpha}_i - \varepsilon \max\limits_{\nu=1,\dots,n} \frac{\sigma_\nu^{(i)}}{\tau_\nu}} =: T_i(A,b,\varepsilon,l_{ij},l_i)$$

Beweis :

Wir gehen aus von Abschätzung (A1) von Satz II.1 .

1. Auswertung von $\max\limits_{R,f} \|A^{-1}(f-Rx)\|_i$:

Nach Hilfssatz II.5 erhält man die folgende Beziehung :

$$\{ f-Rx \mid f \in \mathcal{F},\ R \in \mathcal{R} \} = \{ t = (t_i) \mid |t_i| \le \varepsilon \tau_i \}$$

Nach Hilfssatz II.4 folgt hieraus unmittelbar :

$$\max_{R,f} \|A^{-1}(f-Rx)\|_i = \varepsilon \max_{e \in \mathcal{E}} \|\bar{A}e\|_i = \frac{\varepsilon}{\bar{\alpha}_i}$$

2. Auswertung von $\max\limits_{R \in \mathcal{R}} \|A^{-1}R\|_i = \max\limits_{R,\|s\|_i=1} \|A^{-1}Rs\|_i$

a) Nach Hilfssatz II.4 und Hilfssatz II.5 erhält man :

$$\{ Rs \mid R \in \mathcal{R},\ \|s\|_i = 1 \} \subset \{ t = (t_j) \mid |t_j| \le \varepsilon \sigma_j^{(i)} \}$$

$$\max_{R,\|s\|_i=1} \|A^{-1}Rs\|_i = \max_{e \in \mathcal{E}^*} \|A^* e\|_i = \frac{\varepsilon}{\alpha_i^*}$$

b) Es sei $B = (b_{ij})$ mit $b_{ij} = \begin{cases} 0 & \text{für } i \neq j \\ \tau_i & \text{für } i = j \end{cases}$

Dann gilt : $A^{-1}B = \bar{A}$; $\max\limits_{R,\|s\|_i=1} \|A^{-1}Rs\|_i = \max\limits_{R,\|s\|_i=1} \|\bar{A}B^{-1}Rs\|_i$

Nach Hilfssatz II.4 und Hilfssatz II.5 erhält man :

$$\{ B^{-1}Rs \mid R \in \mathcal{R},\ \|s\|_i = 1 \} \subset \{ t = (t_j) \mid |t_j| \le \varepsilon \frac{\sigma_j^{(i)}}{\tau_j} \}$$

$$\max_{R,\|s\|_i=1} \|A^{-1}Rs\|_i = \max_{R,\|s\|_i=1} \|\bar{A}B^{-1}Rs\|_i \le$$

$$\le \varepsilon \max_{\nu=1,\dots,n} \frac{\sigma_\nu^{(i)}}{\tau_\nu} \max_{e \in \mathcal{E}} \|\bar{A}e\|_i = \frac{\varepsilon \max\limits_{\nu=1,\dots,n} \frac{\sigma_\nu^{(i)}}{\tau_\nu}}{\bar{\alpha}_i}$$

Setzt man die Abschätzungen für $\max\limits_{R,f} \|A^{-1}(f-Rx)\|_i$ und $\max\limits_{R,\|s\|_i=1} \|A^{-1}Rs\|_i$ in die Abschätzung (A1) von Satz II.1 ein, so erhält man nach elementarer Umformung die Abschätzungen S_i und T_i .

Zusatz 1 :

Für die Normen $i = 1,2,3$ lassen sich die Terme $\sigma_\nu^{(i)}$ leicht angeben.

$$\sigma_\nu^{(1)} = \sum_{j=1}^{n} l_{\nu j} \; ; \quad \sigma_\nu^{(2)} = \max_{j=1,\dots,n} l_{\nu j} \; ; \quad \sigma_\nu^{(3)} = \left(\sum_{j=1}^{n} l_{\nu j}^2 \right)^{1/2}$$

Zusatz 2 :

Für die Norm $i = 1$ kann man den Term $\max_{R \in \mathcal{R}} \| A^{-1} R \|_1$ exakt angeben.

Es sei $A^{-1} = (\alpha_{ij})$, dann gilt die Beziehung :

$$\max_{R \in \mathcal{R}} \| A^{-1} R \|_1 = \varepsilon \max_{i=1,\dots,n} \sum_{\lambda=1}^{n} \sum_{j=1}^{n} |\alpha_{i\lambda}| \, l_{j\lambda}$$

Bemerkung :

Die Abschätzung S_i ist besser als T_i ; allerdings wegen der zusätzlichen Berechnung der Größen α_i^* ist die numerische Auswertung von S_i wesentlich aufwendiger.

Legt man die Abschätzung (A2) von Satz II.1 zugrunde, so erhält man für die Normen $i = 1,2,3$ eine weitere Abschätzung für die Größe ρ_i . Diese läßt sich numerisch leichter auswerten, ist allerdings gröber als die eben angegebene Abschätzung.

Satz II.7 :

Gegeben seien das lineare Gleichungssystem $Ax = b$ $\det A \neq 0$ und

$$\mathcal{R} := \left\{ R = (r_{ij}) \,\middle|\, |r_{ij}| \leq \varepsilon l_{ij} \right\} \; ; \; \mathcal{F} := \left\{ f = (f_i) \,\middle|\, |f_i| \leq \varepsilon l_i \right\}$$

$$\tau_\nu := l_\nu + \sum_{j=1}^{n} l_{\nu j} |x_j| \; ; \quad L := (l_{ij}) \, .$$

Die in den folgenden Abschätzungen auftretenden Nenner seien > 0 . Dann gelten für die Normen $i = 1,2,3$ die Abschätzungen :

$$\rho_1 \leq \frac{\varepsilon \max_{\nu=1,\dots,n} \tau_\nu}{\frac{1}{\| A^{-1} \|_1} - \varepsilon \| L \|_1}$$

$$\rho_2 \leq \frac{\varepsilon \sum_{\nu=1}^{n} \tau_\nu}{\frac{1}{\| A^{-1} \|_2} - \varepsilon \| L \|_2}$$

$$\rho_3 \leq \frac{\varepsilon \left(\sum_{\nu=1}^{n} \tau_\nu^2 \right)^{1/2}}{\frac{1}{\|A^{-1}\|_3} - \varepsilon \left(\sum_{i,j=1}^{n} l_{ij}^2 \right)^{1/2}}$$

Beweis:

Wir gehen aus von Abschätzung (A2) von Satz II.1 .

$$\max_{R,f} \|f-Rx\|_1 = \varepsilon \max_{\nu=1,\dots,n} \tau_\nu \quad ; \quad \max_{R \in \mathcal{R}} \|R\|_1 = \varepsilon \|L\|_1$$

$$\max_{R,f} \|f-Rx\|_2 = \varepsilon \sum_{\nu=1}^{n} \tau_\nu \quad ; \quad \max_{R \in \mathcal{R}} \|R\|_2 = \varepsilon \|L\|_2$$

$$\max_{R,f} \|f-Rx\|_3 = \varepsilon \left(\sum_{\nu=1}^{n} \tau_\nu^2 \right)^{1/2} ; \quad \max_{R \in \mathcal{R}} \|R\|_3 \leq \varepsilon \left(\sum_{i,j=1}^{n} l_{ij}^2 \right)^{1/2}$$

Durch Einsetzen in die Abschätzung (A2) erhält man unmittelbar die im Satz angegebenen Abschätzungen.

Wir gehen aus von den folgenden Fehlermatrizen

$$\mathcal{R} := \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon > 0 \right\}$$
$$\mathcal{F} := \left\{ f = (f_j) \mid \|f\|_i \leq \delta > 0 \right\}$$

und werden für die Größe ρ_i nach oben und nach unten Abschätzungen angeben.

Hilfssatz II.6 :

Zu den Mengen $\mathcal{R} := \left\{ R \mid \|R\|_i \leq \varepsilon \geq 0 \right\}$; $\mathcal{F} := \left\{ f \mid \|f\|_i \leq \delta \geq 0 \right\}$ existieren Zahlen $\varepsilon_o^{(i)} \geq 0$, $\delta_o^{(i)} \geq 0$ mit den Eigenschaften :

$$\left\{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon_o^{(i)} \right\} \subset \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon \right\}$$
$$\left\{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon' \right\} \not\subset \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon \right\} \quad \text{für } \varepsilon' > \varepsilon_o^{(i)}$$

$$\left\{ f = (f_i) \mid |f_i| \leq \delta_o^{(i)} \right\} \subset \left\{ f = (f_i) \mid \|f\|_i \leq \delta \right\}$$
$$\left\{ f = (f_i) \mid |f_i| \leq \delta' \right\} \not\subset \left\{ f = (f_i) \mid \|f\|_i \leq \delta \right\} \quad \text{für } \delta' > \delta_o^{(i)}$$

Beweis :

Die Aussage ist wegen der Kompaktheit der auftretenden Mengen unmittelbar klar.

Zusatz 1 :

Für die Normen $i = 1,2,3$ lassen sich die Terme $\sigma_\nu^{(i)}$ leicht angeben.

$$\sigma_\nu^{(1)} = \sum_{j=1}^{n} l_{\nu j} \; ; \quad \sigma_\nu^{(2)} = \max_{j=1,\dots,n} l_{\nu j} \; ; \quad \sigma_\nu^{(3)} = \left(\sum_{j=1}^{n} l_{\nu j}^2 \right)^{1/2}$$

Zusatz 2 :

Für die Norm $i = 1$ kann man den Term $\max_{R \in \mathcal{R}} \| A^{-1} R \|_1$ exakt angeben.

Es sei $A^{-1} = (\alpha_{ij})$, dann gilt die Beziehung :

$$\max_{R \in \mathcal{R}} \| A^{-1} R \|_1 = \varepsilon \max_{i=1,\dots,n} \sum_{\lambda=1}^{n} \sum_{j=1}^{n} |\alpha_{ij}| \, l_{j\lambda}$$

Bemerkung :

Die Abschätzung S_i ist besser als T_i ; allerdings wegen der zusätzlichen Berechnung der Größen α_i^* ist die numerische Auswertung von S_i wesentlich aufwendiger.

Legt man die Abschätzung (A2) von Satz II.1 zugrunde, so erhält man für die Normen $i = 1,2,3$ eine weitere Abschätzung für die Größe ρ_i . Diese läßt sich numerisch leichter auswerten, ist allerdings gröber als die eben angegebene Abschätzung.

Satz II.7 :

Gegeben seien das lineare Gleichungssystem $Ax = b$ $\det A \neq 0$ und

$$\mathcal{R} := \left\{ R = (r_{ij}) \,\middle|\, |r_{ij}| \leq \varepsilon l_{ij} \right\} \; ; \; \mathcal{F} := \left\{ f = (f_i) \,\middle|\, |f_i| \leq \varepsilon l_i \right\}$$

$$\tau_\nu := l_\nu + \sum_{j=1}^{n} l_{\nu j} |x_j| \; ; \quad L := (l_{ij}) \, .$$

Die in den folgenden Abschätzungen auftretenden Nenner seien > 0 . Dann gelten für die Normen $i = 1,2,3$ die Abschätzungen :

$$\rho_1 \leq \frac{\varepsilon \max_{\nu=1,\dots,n} \tau_\nu}{\frac{1}{\| A^{-1} \|_1} - \varepsilon \| L \|_1}$$

$$\rho_2 \leq \frac{\varepsilon \sum_{\nu=1}^{n} \tau_\nu}{\frac{1}{\| A^{-1} \|_2} - \varepsilon \| L \|_2}$$

$$\rho_3 \leq \frac{\varepsilon \left(\sum_{\nu=1}^{n} \tau_\nu^2 \right)^{1/2}}{\frac{1}{\|A^{-1}\|_3} - \varepsilon \left(\sum_{i,j=1}^{n} l_{ij}^2 \right)^{1/2}}$$

Beweis:

Wir gehen aus von Abschätzung (A2) von Satz II.1 .

$$\max_{R,f} \|f-Rx\|_1 = \varepsilon \max_{\nu=1,\dots,n} \tau_\nu \quad ; \quad \max_{R \in \mathfrak{R}} \|R\|_1 = \varepsilon \|L\|_1$$

$$\max_{R,f} \|f-Rx\|_2 = \varepsilon \sum_{\nu=1}^{n} \tau_\nu \quad ; \quad \max_{R \in \mathfrak{R}} \|R\|_2 = \varepsilon \|L\|_2$$

$$\max_{R,f} \|f-Rx\|_3 = \varepsilon \left(\sum_{\nu=1}^{n} \tau_\nu^2 \right)^{1/2} \quad ; \quad \max_{R \in \mathfrak{R}} \|R\|_3 \leq \varepsilon \left(\sum_{i,j=1}^{n} l_{ij}^2 \right)^{1/2}$$

Durch Einsetzen in die Abschätzung (A2) erhält man unmittelbar die im Satz angegebenen Abschätzungen.

Wir gehen aus von den folgenden Fehlermatrizen

$$\mathfrak{R} := \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon > 0 \right\}$$
$$\mathfrak{F} := \left\{ f = (f_j) \mid \|f\|_i \leq \delta > 0 \right\}$$

und werden für die Größe ρ_i nach oben und nach unten Abschätzungen angeben.

Hilfssatz II.6 :

Zu den Mengen $\mathfrak{R} := \left\{ R \mid \|R\|_i \leq \varepsilon \geq 0 \right\}$; $\mathfrak{F} := \left\{ f \mid \|f\|_i \leq \delta \geq 0 \right\}$ existieren Zahlen $\varepsilon_o^{(i)} \geq 0$, $\delta_o^{(i)} \geq 0$ mit den Eigenschaften :

$$\left\{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon_o^{(i)} \right\} \subset \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon \right\}$$
$$\left\{ R = (r_{ij}) \mid |r_{ij}| \leq \varepsilon' \right\} \not\subset \left\{ R = (r_{ij}) \mid \|R\|_i \leq \varepsilon \right\} \quad \text{für } \varepsilon' > \varepsilon_o^{(i)}$$

$$\left\{ f = (f_i) \mid |f_i| \leq \delta_o^{(i)} \right\} \subset \left\{ f = (f_i) \mid \|f\|_i \leq \delta \right\}$$
$$\left\{ f = (f_i) \mid |f_i| \leq \delta' \right\} \not\subset \left\{ f = (f_i) \mid \|f\|_i \leq \delta \right\} \quad \text{für } \delta' > \delta_o^{(i)}$$

Beweis :

Die Aussage ist wegen der Kompaktheit der auftretenden Mengen unmittelbar klar.

Zusatz :

Für die Normen $i = 1,2,3$ lassen sich die Größen $\varepsilon_o^{(i)}$ und $\delta_o^{(i)}$ unmittelbar angeben. Es gilt :

$$\varepsilon_o^{(1)} = \frac{\varepsilon}{n} \; ; \quad \varepsilon_o^{(2)} = \frac{\varepsilon}{n} \; ; \quad \varepsilon_o^{(3)} = \frac{\varepsilon}{n}$$

$$\delta_o^{(1)} = \delta \; ; \quad \delta_o^{(2)} = \frac{\delta}{n} \; ; \quad \delta_o^{(3)} = \frac{\delta}{\sqrt{n}}$$

Satz II.8 :

Gegeben seien das lineare Gleichungssystem $Ax = b$ $\det A \neq 0$ und

$$\mathcal{R} := \left\{ R \mid \|R\|_i \leq \varepsilon \right\} \; ; \quad \mathcal{F} := \left\{ f \mid \|f\|_i \leq \delta \right\}$$

Es sei $\varepsilon \|A^{-1}\|_i < 1$. Dann gelten die folgenden Abschätzungen :

$$\underline{M}_i(A,b,\varepsilon,\delta) := \frac{\varepsilon_o^{(i)} \|x\|_2 + \delta_o^{(i)}}{\alpha_i + \varepsilon_o^{(i)} \max\limits_{\|s\|_i = 1} \|s\|_2} \leq \rho_i$$

$$\rho_i \leq \frac{\varepsilon \|x\|_i + \delta}{\frac{1}{\|A^{-1}\|_i} - \varepsilon} =: \overline{M}_i(A,b,\varepsilon,\delta)$$

Für die Normen $i = 1,2,3$ gilt insbesondere :

$$\underline{M}_1 := \frac{\varepsilon \frac{\|x\|_2}{n} + \delta}{\alpha_1 + \varepsilon} \leq \rho_1 \leq \frac{\varepsilon \|x\|_1 + \delta}{\frac{1}{\|A^{-1}\|_1} - \varepsilon} =: \overline{M}_1$$

$$\underline{M}_2 := \frac{\varepsilon \|x\|_2 + \delta}{n \cdot \alpha_2 + \varepsilon} \leq \rho_2 \leq \frac{\varepsilon \|x\|_2 + \delta}{\frac{1}{\|A^{-1}\|_2} - \varepsilon} =: \overline{M}_2$$

$$\underline{M}_3 := \frac{\varepsilon \frac{\|x\|_2}{\sqrt{n}} + \delta}{\sqrt{n}\, \alpha_3 + \varepsilon} \leq \rho_3 \leq \frac{\varepsilon \|x\|_3 + \delta}{\frac{1}{\|A^{-1}\|_3} - \varepsilon} =: \overline{M}_3$$

Beweis :

I. $\underline{M}_i(A,b,\varepsilon,\delta) \leq \rho_i$

Wegen Hilfssatz II.6 genügt die folgende Festsetzung :

$$\underline{M}_i(A,b,\varepsilon,\delta) := \underline{\phi}_i(A,b,\varepsilon_o^{(i)},\delta_o^{(i)})$$

Durch Auswertung dieser Gleichung für die Normen $i = 1,2,3$ erhält man die Abschätzungen $\underline{M}_1$, $\underline{M}_2$, $\underline{M}_3$.

II. $\rho_i \leq \bar{M}_i(A,b,\varepsilon,\delta)$

Wir gehen aus von Abschätzung (A4) von Satz II.1 . Beachtet man $\max_{R\in\mathcal{R}} \|R\|_i = \varepsilon$ und $\max_{f\in\mathcal{F}} \|f\|_i = \delta$, so erhält man durch Einsetzen unmittelbar die angegebene Abschätzung.

Bemerkung :

Zu den Abschätzungen $\bar{M}_i$ vergleiche man die Literaturangaben [1] , [2] , [7] , [8] .

§ 2. Abhängigkeit der Lösung eines linearen Gleichungssystems von zufälligen Fehlern in den Ausgangsdaten.

Wir gehen aus von dem numerisch gegebenen linearen Gleichungssystem

$$Ax = b \quad \det A \neq 0 \quad x \text{ explizit bekannt.}$$

Eine Änderung der Ausgangsmatrix bewirkt eine Änderung der Lösung.

$$\begin{array}{l} A \longrightarrow A+R \\ x \longrightarrow x+\delta x \end{array} \quad \text{mit} \quad \det(A+R) \neq 0$$

Dabei gilt für den Defektvektor δx die folgende Beziehung :

$$\delta x = -(A+R)^{-1}Rx$$

Wir interessieren uns für die Abhängigkeit des Defektvektors δx von zufälligen Fehlern in der Koeffizientenmatrix.

Die genaue Formulierung des Problems geschieht in dem folgenden wahrscheinlichkeitstheoretischen Modell :

Gegeben sei der Wahrscheinlichkeitsraum

$$\Omega = (\mathcal{R}, \gamma, W)$$

$\mathcal{R}$ = Menge aller Matrizen $R = (r_{ij})$

γ = σ -Algebra der Borelmengen

W = Wahrscheinlichkeitsmaß

Wir wollen für das Wahrscheinlichkeitsmaß W die folgende Zusatzvoraussetzung machen : $W(\{ R \mid \det(A+R) = 0 \}) = 0$

Die folgende zufällige Variable ist demnach höchstens auf einer Menge vom Maße Null nicht definiert :

$$\xi^{(i)}: \Omega \ni R \longrightarrow \|\delta x\|_i = \|(A+R)^{-1}Rx\|_i$$

Problem:

Zu vorgegebener Sicherheitswahrscheinlichkeit $1/2 < p = 1 - \varepsilon$ ist der Quantil γ der Ordnung p gesucht. Man beachte, daß der nicht notwendige eindeutige Wert γ der folgenden Bedingung genügt:

$$W(\{ \xi^{(i)} \leq \gamma \}) \geq p$$

Für die weiteren Betrachtungen wählen wir, falls der Quantil nicht eindeutig ist, stets das Minimum.

Die Verteilung der zufälligen Variablen $\xi^{(i)}$ ist selbst für ein einfaches Wahrscheinlichkeitsmaß W so kompliziert, daß eine exakte Berechnung der Größe γ nicht möglich ist. Wir werden für γ eine Abschätzung angeben. Diese Abschätzung ist allerdings nur für kleine Fehler in den Ausgangsdaten gut. Aber auch die numerische Auswertung der Abschätzung bereitet im allgemeinen große Schwierigkeiten. Sie gelingt hier nur für ein sehr spezielles Wahrscheinlichkeitsmaß W und für einen sehr speziellen Typ von Koeffizientenmatrizen sowie für die Normen $i = 1,2$.

Nach Satz I.12 gilt für $\|A^{-1}R\|_i < 1$ die folgende Abschätzung:

$$\|\delta x\|_i \leq \frac{\|A^{-1}Rx\|_i}{1 - \|A^{-1}R\|_i}$$

Diese Beziehung macht es naheliegend, die Untersuchung der zufälligen Variablen $\xi^{(i)}$ auf die Untersuchung der beiden folgenden zufälligen Variablen zurückzuführen :

$$\xi_1^{(i)}: \Omega \ni R \longrightarrow \|A^{-1}R\|_i$$

$$\xi_2^{(i)}: \Omega \ni R \longrightarrow \|A^{-1}Rx\|_i$$

Hilfssatz II.7 :

ξ_1, ξ_2 seien zwei unabhängige zufällige Variable. $0 \leq \delta < 1$

Für ein beliebiges μ gelte die folgende Beziehung :

$$W(\{\xi_1 < \mu\}) \geq W(\{\xi_2 \leq \mu\}) - \delta$$

Zu gegebener Wahrscheinlichkeit $p < 1 - \delta$ seien λ und λ^* der (minimale) Quantil der Ordnung p bzw. $p+\delta$. Das heißt :

$$W(\{\xi_1 \leq \lambda\}) \geq p \; ; \; W(\{\xi_2 \leq \lambda^*\}) \geq p+\delta$$

Dann gilt die folgende Abschätzung : $\lambda^* \geq \lambda$

Beweis :

Annahme : $\lambda^* < \lambda$

Dann gilt : $W(\{\xi_1 \leq \lambda^*\}) < p \; ; \; W(\{\xi_2 \leq \lambda^*\}) \geq p+\delta$

Daraus folgt : $W(\{\xi_1 \leq \lambda^*\}) < W(\{\xi_2 \leq \lambda^*\}) - \delta$

Dies ist ein Widerspruch zur Voraussetzung.

Wir sind jetzt in der Lage, unter bestimmten Voraussetzungen mit Hilfe der zufälligen Variablen $\xi_1^{(i)}$ und $\xi_2^{(i)}$ eine Abschätzung für die Größe γ anzugeben.

Satz II.9 :

Gegeben sei die Sicherheitswahrscheinlichkeit $p = 1 - \epsilon < 1$. Es existiere ein Wert α mit $0 < \alpha < 1$ und ein Wert $0 \leq \delta < \epsilon$ mit

$$W(\{\xi_1^{(i)} \leq \alpha\}) = 1 - \delta$$

γ, γ^* seien die (minimalen) Quantile der Ordnung p bzw. $p+\delta$ bezüglich $\xi^{(i)}$ bzw. $\frac{\xi_2^{(i)}}{1-\alpha}$, d.h. es gelten die beiden Relationen :

$$W(\{\xi^{(i)} \leq \gamma\}) \geq p \; ; \; W(\{\frac{\xi_2^{(i)}}{1-\alpha} \leq \gamma^*\}) \geq p+\delta \; ,$$

dann gilt die folgende Abschätzung : $\gamma^* \geq \gamma$

Beweis :

$\mathfrak{R} \supset \mathfrak{R}_c := \{R \mid \|A^{-1}R\|_i \leq \alpha\}$. Dann gilt : $W(\mathfrak{R}_c) = 1 - \delta$.

Für ein beliebiges μ gelten die folgenden Relationen:

$$W(\{\|\delta x\|_i \leq \mu\}) \geq W(\{R \mid \frac{\|A^{-1}Rx\|_i}{1 - \|A^{-1}R\|_i} \leq \mu\}) =$$

$= W\left(\left\{ R \in \mathcal{R}_0 \;\middle|\; \frac{\|A^{-1}Rx\|_i}{1-\|A^{-1}R\|_i} \le \mu \right\}\right) + W\left(\left\{ R \notin \mathcal{R}_0 \;\middle|\; \frac{\|A^{-1}Rx\|_i}{1-\|A^{-1}R\|_i} \le \mu \right\}\right) \ge$

$\ge W\left(\left\{ R \in \mathcal{R}_0 \;\middle|\; \frac{\|A^{-1}Rx\|_i}{1-\alpha} \le \mu \right\}\right) + W\left(\left\{ R \notin \mathcal{R}_0 \;\middle|\; \frac{\|A^{-1}Rx\|_i}{1-\alpha} \le \mu \right\}\right) - \delta =$

$= W\left(\left\{ R \;\middle|\; \frac{\|A^{-1}Rx\|_i}{1-\alpha} \le \mu \right\}\right) - \delta$. Daraus folgt die Ungleichung:

$$W\left(\left\{ \xi^{(i)} \le \mu \right\}\right) \ge W\left(\left\{ \frac{\xi_2^{(i)}}{1-\alpha} \le \mu \right\}\right) - \delta$$

Nach Hilfssatz II.7 folgt hieraus unmittelbar die Behauptung des Satzes.

Bemerkung :

1. Es brauchen natürlich keine Zahlen $\delta < \varepsilon, \alpha < 1$ mit der angegebenen Eigenschaft zu existieren. Im Falle der Existenz sind sie im allgemeinen nicht eindeutig. In diesem Fall wähle man α, δ so, daß γ^* möglichst klein wird.

2. Für die Auswertung der Abschätzung γ^* braucht man nur noch die wesentlich einfacheren zufälligen Variablen $\xi_1^{(i)}$ und $\xi_2^{(i)}$ zu untersuchen.

3. Die Bedingungen $\delta < \varepsilon, \alpha < 1$ sind erfüllt, wenn die Fehler in der Koeffizientenmatrix nicht zu groß sind.

4. Die in dem Satz angegebene Abschätzung ist umso besser je kleiner die Größen α und δ sind. Sehr gut ist die Abschätzung für $\alpha << 1$ und $\delta << \varepsilon$.

Die weiteren Ausführungen gelten der Untersuchung der zufälligen Variablen $\xi_1^{(i)}$ und $\xi_2^{(i)}$. Dabei werden wir uns auf das folgende Wahrscheinlichkeitsmaß W beschränken: W sei dadurch festgelegt, daß die Koeffizienten r_{ij} unabhängige nach $N(0, \sigma^2)$ verteilte zufällige Variable sind.

Wichtig für die Diskussion der zufälligen Variablen $\xi_1^{(i)}$ und $\xi_2^{(i)}$ ist das Verhalten der folgenden zufälligen Matrix η :

$$\eta : \Omega \ni R \longrightarrow A^{-1}R = : S = (s_{ij})$$

Nach Satz I.13 sind die Koeffizienten s_{ij} zufällige Variable,

welche nach $N(0, \sigma_{ij}^2)$ verteilt sind. Dabei gilt :

$$\sigma_{ij} = \sigma \left(\sum_{\lambda=1}^{n} \alpha_{i\lambda}^2 \right)^{1/2}$$

Für die numerische Auswertung der zufälligen Variablen $\xi_1^{(i)}$ und $\xi_2^{(i)}$ erweist es sich als günstig, wenn die zufälligen Variablen unabhängig sind. Wegen der folgenden Beziehung

$$s_{ij} = \sum_{\lambda=1}^{n} \alpha_{i\lambda} r_{\lambda j}$$

sind die Koeffizienten der selben Spalte von S nach Satz I.16 nur dann exakt unabhängig, wenn die Zeilenvektoren von A^{-1} paarweise aufeinander senkrecht stehen. Nach Satz I.17 können die zufälligen Variablen s_{ij} auch noch dann als "praktisch" unabhängig angesehen werden, wenn für die Zeilenvektoren $z_1, z_2, \ldots, z_n$ von A^{-1} die folgende Bedingung erfüllt ist:

$$\frac{|z_i^T z_j|}{\|z_i\|_3 \, \|z_j\|_3} \leq \gamma \quad \text{für} \quad i \neq j \quad \text{mit} \quad 0 \leq \gamma \ll 1$$

Über die erforderliche Größenordnung der Zahl γ werden hier keine weiteren Betrachtungen angestellt.

Bei den weiteren Untersuchungen gehen wir stets davon aus, daß wir eine Koeffizientenmatrix dieser Art vorliegen haben.

Bemerkung :

Wir denken uns die Matrix A als Realisation einer zufälligen Matrix. Zu vorgegebenem $0 < \gamma < 1$ ist die Wahrscheinlichkeit, daß die Zeilenvektoren der Matrix der angegebenen Bedingung genügen um so größer, je größer die Ordnung der Matrix ist.

Nach diesen Vorbereitungen sind wir in der Lage, die Aussage von Satz II.9 für die Normen $i = 1,2$ numerisch auszuwerten. Gegeben seien für die weiteren Ausführungen eine Sicherheitswahrscheinlichkeit $p = 1 - \varepsilon$ und $0 < \varepsilon < 1/2$.

I. Diskussion der zufälligen Variablen $\xi_1^{(i)}$:

1. $i = 1$

Wir erhalten $\xi_1^{(1)} = \max\limits_{i=1,\ldots,n} \sum\limits_{j=1}^{n} |s_{ij}|$

Für die zufälligen Variablen $\eta_i = \sum_{j=1}^{n} |s_{ij}|$ gilt nach Satz I.13 und Satz I.14 für den Mittelwert μ_i^* und die Streuung σ_i^*:

$$\mu_i^* = \frac{2\sigma n}{\sqrt{2\pi}} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

$$\sigma_i^* = \sigma \sqrt{n} \left(1 - \frac{2}{\pi}\right)^{1/2} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

Für hinreichen große n und unter geeigneten Voraussetzungen für die Größen σ_{ij} sind die zufälligen Variablen η_i näherungsweise verteilt nach $N(\mu_i^*, \sigma_i^{*2})$.

Wir suchen Zahlen $0 < \alpha < 1$ und $\delta < \varepsilon$ mit der Eigenschaft:

$$W\{ \xi_1^{(1)} \leq \alpha \} = 1 - \delta$$

Nach Satz I.15 empfiehlt sich der folgende Ansatz:

$$\alpha = \max_{i=1,\dots,n} \mu_i^* + c \max_{i=1,\dots,n} \sigma_i^*$$

Die Größe $c = c(\sqrt[n]{1-\delta})$ läßt sich einer Tabelle über Normalverteilung entnehmen.

Man beachte, daß für hinreichend große n im allgemeinen die Beziehung $\frac{\sigma_i^*}{\mu_i^*} << 1$ gilt. Daraus folgt die Näherung:

$$\alpha \approx \max_{i=1,\dots,n} \mu_i^*$$

2. $i = 2$

Wir erhalten $\xi_1^{(2)} = \max_{j=1,\dots,n} \sum_{i=1}^{n} |s_{ij}|$

Für die zufälligen Variablen $\eta_j = \sum_{i=1}^{n} |s_{ij}|$ gilt nach Satz I.13 und Satz I.14 für den Mittelwert μ_j^* und die Streuung σ_j^*:

$$\mu_j^* = \frac{2\sigma}{\sqrt{2\pi}} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

$$\sigma_j^* = \sigma \left(1 - \frac{2}{\pi}\right)^{1/2} \left(\sum_{i,j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

Insbesondere haben die zufälligen Variablen η_j alle die selbe Verteilung.

Für hinreichend große n und unter geeigneten Voraussetzungen für die Größen σ_{ij} sind die zufälligen Variablen η_j näherungsweise verteilt nach $N(\mu_j^*, \sigma_j^{*2})$.

Wir suchen Zahlen $0 < \alpha < 1$ und $\delta < \varepsilon$ mit der Eigenschaft :

$$W(\{\xi_1^{(2)} \leq \alpha\}) = 1-\delta$$

Nach Satz I.15 empfiehlt sich der folgende Ansatz :

$$\alpha = \mu_j^* + c\,\sigma_j^*$$

Die Größe $c = c(\sqrt[n]{1-\delta})$ läßt sich einer Tabelle über Normalverteilung entnehmen.

Man beachte, daß für hinreichend große n im allgemeinen die Beziehung $\frac{\sigma_j^*}{\mu_j^*} << 1$ gilt. Daraus folgt die Näherung :

$$\alpha \approx \mu_j^*$$

II. Diskussion der zufälligen Variablen $\xi_2^{(i)}$:

Die Komponenten ω_ν des Vektors $A^{-1}Rx$ sind nach Satz I.13 unabhängige, nach $N(0,\sigma_\nu^2)$ verteilte zufällige Variable mit

$$\sigma_\nu = \sigma\,(x^T x)^{1/2}\left(\sum_{j=1}^{n} \alpha_{\nu j}^2\right)^{1/2}$$

1. $i = 1$

Wir erhalten $\xi_2^{(1)} = \max\limits_{\nu=1,\dots,n} |\omega_\nu|$

Gesucht ist der Quantil s^* der Ordnung $p+\delta$. Es gilt :

$$W(\{\xi_2^{(1)} \leq s^*\}) = p+\delta$$

Nach Satz I.15 erhält man die folgende Abschätzung :

$$s^* \leq c \max_{\nu=1,\dots,n} \sigma_\nu = c\,\sigma\sqrt{x^T x}\,\max_{i=1,\dots,n}\left(\sum_{j=1}^{n} \alpha_{ij}^2\right)^{1/2}$$

Die Größe $c = c(\sqrt[n]{p+\delta})$ läßt sich einer Tabelle über Normalverteilung entnehmen.

<u>Ergebnis :</u>

Wir setzen $\gamma^* := \frac{s^*}{1-\alpha}$ und erhalten damit die gesuchte Abschätzung für γ.

2. $i = 2$

Wir erhalten $\xi_2^{(2)} = \sum\limits_{\nu=1}^{n} |\omega_\nu|$

Nach Satz I.13 und Satz I.14 erhält man für den Mittelwert μ^* und die Streuung σ^* von $\xi_2^{(2)}$:

$$\mu^* = \frac{2\sigma}{\sqrt{2\pi}} (x^T x)^{1/2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

$$\sigma^* = \sigma \left(1 - \frac{2}{\pi}\right)^{1/2} (x^T x)^{1/2} \left(\sum_{i,j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$$

Für hinreichend große n und unter geeigneten Voraussetzungen für die Größen μ_ν, σ_ν ist $\xi_2^{(2)}$ näherungsweise verteilt nach $N(\mu^*, \sigma^{2*})$.

Gesucht ist der Quantil s^* der Ordnung $p+\delta$. Es gilt :

$$W(\{ \xi_2^{(2)} \leq s^* \}) = p + \delta$$

Nach Satz I.15 erhält man die folgende Abschätzung :

$$s^* \leq \mu^* + c\, \sigma^*$$

Die Größe $c = c(\sqrt[n]{p+\delta})$ läßt sich einer Tabelle über Normalverteilung entnehmen.

Ergebnis :

Wir setzen $\gamma^* := \frac{s^*}{1-\alpha}$ und erhalten damit die gesuchte Abschätzung für γ.

§ 3. Experimentelle Untersuchungen für zufällige Fehler in den Ausgangsdaten.

In diesem Paragraphen soll der Einfluß von zufälligen Fehlern in der Koeffizientenmatrix eines linearen Gleichungssystems auf die Lösung experimentell untersucht werden. Dabei wird einmal die Ordnung der Koeffizientenmatrix und zum anderen die Größe der Fehler variiert. Die Untersuchungen, die hier angestellt werden, sind so allgemein, daß es in dieser Arbeit nicht gelungen ist, die Ergebnisse durch ein theoretisches Modell zu untermauern.

Die folgenden Ausführungen gelten einer genauen Beschreibung der durchgeführten Experimente.

Wir gehen aus von dem linearen Gleichungssystem $Ax = b$. Die Koeffizienten von A und die komponenten des Vektors b werden als Realisationen einer symmetrischen Gleichverteilung im Intervall $[-M,M]$ gewonnen. Durch Überlagerung der Matrix mit einer Fehler-

matrix R, deren Koeffizienten als Realisationen einer Gleichverteilung im Intervall $[-\varepsilon,\varepsilon]$ gewonnen werden, erhält man das benachbarte Gleichungssystem $(A+R)y = b$.

Wir interessieren uns nun für die relative Abweichung $\frac{\|y-x\|_i}{\|x\|_i}$ bezüglich der Normen $i = 1,2,3$.

Dabei wird sowohl die Ordnung des linearen Gleichungssystems als auch der Quotient $\frac{\varepsilon}{M}$ variiert.

Wir geben jetzt eine Reihe von Zahlenbeispielen an, die wir anschließend kurz interpretieren werden. Wir weisen nachdrücklich darauf hin, daß das angegebene Zahlenmaterial bei weitem nicht ausreicht, um zuverlässige statistische Aussagen zu erhalten.

1. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 10$.

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0392	0,1389	0,7538
i = 2	0,0368	0,1547	1,2491
i = 3	0,0367	0,1503	1,0322

2. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 20$.

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,8460	0,5071	1,0016
i = 2	0,8203	0,5302	0,9371
i = 3	0,8177	0,5296	0,9533

3. Beispiel :

Die Ordnung der verwendeten Matrix n = 50 .

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,4237	0,7005	1,0622
i = 2	0,4278	0,7205	1,0377
i = 3	0,4300	0,7219	1,0230

4. Beispiel :

Die Ordnung der verwendeten Matrix ist n = 10 .

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0211	0,4894	2,9124
i = 2	0,0226	0,5038	2,2633
i = 3	0,0205	0,4981	2,1965

5. Beispiel :

Die Ordnung der verwendeten Matrix ist n = 20 .

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0261	0,6141	1,9728
i = 2	0,0264	0,6066	2,1079
i = 3	0,0273	0,6045	2,0170

6. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 50$.

	$\frac{\varepsilon}{M} = 0{,}01$	$\frac{\varepsilon}{M} = 0{,}1$	$\frac{\varepsilon}{M} = 0{,}5$
i = 1	0,2126	27,7266	1,4199
i = 2	0,2014	25,8046	1,2823
i = 3	0,1940	24,8963	1,2994

7. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 10$.

	$\frac{\varepsilon}{M} = 0{,}01$	$\frac{\varepsilon}{M} = 0{,}1$	$\frac{\varepsilon}{M} = 0{,}5$
i = 1	0,0152	0,1767	1,1332
i = 2	0,0187	0,2227	1,1253
i = 3	0,0185	0,2028	1,0659

8. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 20$.

	$\frac{\varepsilon}{M} = 0{,}01$	$\frac{\varepsilon}{M} = 0{,}1$	$\frac{\varepsilon}{M} = 0{,}5$
i = 1	0,0229	0,2496	0,9522
i = 2	0,0284	0,2859	0,9054
i = 3	0,0275	0,2656	0,9392

9. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 50$.

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0988	0,5947	2,124
i = 2	0,0875	0,5181	1,6529
i = 3	0,0909	0,5228	1,7457

10. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 10$.

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0138	0,7774	0,9641
i = 2	0,0159	0,6844	0,6782
i = 3	0,0145	0,7199	0,7726

11. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 20$.

	$\frac{\varepsilon}{M} = 0,01$	$\frac{\varepsilon}{M} = 0,1$	$\frac{\varepsilon}{M} = 0,5$
i = 1	0,0397	0,5275	2,2455
i = 2	0,0532	0,6117	3,4751
i = 3	0,0508	0,5956	3,0958

12. Beispiel :

Die Ordnung der verwendeten Matrix ist $n = 50$.

	$\frac{\varepsilon}{M} = 0{,}01$	$\frac{\varepsilon}{M} = 0{,}1$	$\frac{\varepsilon}{M} = 0{,}5$
$i = 1$	0,1597	13,5086	0,8387
$i = 2$	0,1452	11,9252	0,9310
$i = 3$	0,1466	12,0916	0,9174

Interpretation :

1. Die relative Abweichung ist für sehr kleine Fehler ($\frac{\varepsilon}{M} = 0{,}01$) klein und bei fester Ordnung der Koeffizientenmatrix nur geringen zufälligen Schwankungen unterworfen. Mit wachsender Ordnung der Koeffizientenmatrix wächst die relative Abweichung.
2. Für größere Fehler ($\frac{\varepsilon}{M} = 0{,}1$ bzw. $\frac{\varepsilon}{M} = 0{,}5$) fällt vor allem die große zufällige Schwankung des relativen Fehlers auf. Auch hier nimmt der relative Fehler mit wachsender Ordnung der Koeffizientenmatrix zu.

§ 4. Stabilität von Matrizen. Konditionszahlen

Die bisherigen Untersuchungen stehen in einem engen Zusammenhang mit dem Problem der Stabilität von linearen Gleichungssystemen. Wir wollen ein Problem stabil nennen, wenn kleine Änderungen in den Ausgangsdaten kleine Änderungen in der Lösung bewirken. Ansonsten heißt das Problem instabil. Die folgenden Ausführungen gelten der Untersuchung der Stabilität der Matrizeninversion. Wir werden zeigen, daß ein enger Zusammenhang zwischen der Stabilität von linearen Gleichungssystemen und der Stabilität der zugehörigen Koeffizientenmatrix besteht. Bei Matrizen spricht man auch von der Kondition und unterscheidet demnach zwischen gut und schlecht konditionierten Matrizen.

Den Zusammenhang zwischen der Stabilität von linearen Gleichungssystemen und der Stabilität der Inversion der zugehörigen Koeffizientenmatrix zeigt der folgende Sachverhalt :

Gegeben sei das lineare Gleichungssystem $Ax = b \quad \det A \neq 0$

Eine Änderung der Ausgangsdaten bewirkt eine Änderung der Lösung.

$$\left.\begin{array}{l} A \longrightarrow A+R \quad \det(A+R) \neq 0 \\ b \longrightarrow b+f \end{array}\right\} \text{ Daraus folgt : } x \longrightarrow x+\delta x$$

Dabei gilt die Beziehung : $\|\delta x\| \leq \|(A+R)^{-1}\| \, \|f-Rx\|$

Daraus sieht man den Zusammenhang zwischen der Stabilität eines linearen Gleichungssystems und der Stabilität der Matrizeninversion. In den folgenden Ausführungen untersuchen wir die Stabilität von Matrizen. Man vergleiche hierzu die Ausführungen in dem Lehrbuch Faddajew [3].

Gegeben sei die Matrix $A = (a_{ij})$ $\det A \neq 0$ und $A^{-1} = (\alpha_{ij})$. Ersetzt man den Koeffizienten a_{ij} durch $a'_{ij} = a_{ij} + \varepsilon$, so geht der Koeffizient $\alpha_{ji} \neq 0$ über in $\alpha'_{ji} = \dfrac{1}{\frac{1}{\alpha_{ji}} + \varepsilon}$.

Folgerung :

Ist α_{ij} ein betragsmäßig großer Koeffizient von A^{-1}, so bewirkt eine kleine Änderung des Koeffizienten a_{ji} eine große Änderung von α_{ij}. Große Elemente in der inversen Matrix bedeuten also Instabilität der Matrix.

Diese Aussage wird durch eine weitere Überlegung bestätigt.

Es sei $E_{ij} = (e_{kl})$ mit $e_{kl} = \begin{cases} 0 & \text{für } (k,l) \neq (i,j) \\ 1 & \text{für } (k,l) = (i,j) \end{cases}$

$E = A^{-1}A$ Daraus folgt : $\dfrac{\partial (A^{-1}A)}{\partial a_{ij}} = 0 = \dfrac{\partial A}{\partial a_{ij}} A^{-1} + A \dfrac{\partial A^{-1}}{\partial a_{ij}}$

Daraus folgt : $\dfrac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1} \dfrac{\partial A}{\partial a_{ij}} A^{-1} = -A^{-1} E_{ij} A^{-1} = -(A^{-1}E_{i_1})(A^{-1}E_{1j}) =$

$$= -\begin{pmatrix} \alpha_{1i} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \alpha_{ni} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \alpha_{j1} & \cdots & \alpha_{jn} \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} = -\begin{pmatrix} \alpha_{1i}\alpha_{j1} & \cdots & \alpha_{1i}\alpha_{jn} \\ \vdots & & \vdots \\ \alpha_{ni}\alpha_{j1} & \cdots & \alpha_{ni}\alpha_{jn} \end{pmatrix}$$

Daraus folgt $\frac{\partial \alpha_{kl}}{\partial a_{ij}} = -\alpha_{ki}\alpha_{jl}$; $d\alpha_{kl} = -\sum_{i,j=1}^{n} \alpha_{ki}\alpha_{jl}\, da_{ij}$

Große Elemente von A^{-1} bewirken also starke Empfindlichkeit gegenüber Änderungen der Koeffizienten von A und damit Instabilität.

Satz :

$x = A^{-1}b$. Daraus folgt: $\frac{\partial x}{\partial a_{ij}} = \frac{\partial A^{-1}}{\partial a_{ij}} b = A^{-1}E_{ij}A^{-1}b = -A^{-1}E_{ij}x$

Daraus folgt: $\frac{\partial x}{\partial a_{ij}} = -x_j \begin{pmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ni} \end{pmatrix}$

Große Elemente von A^{-1} haben also im allgemeinen Instabilität des linearen Gleichungssystems zur Folge.

In der Literatur (vgl. [3] , [4] , [6] , [7] , [8]) werden zur Beschreibung der Stabilität von Matrizen eine Reihe von Konditionszahlen (K) vorgeschlagen. Die Konditionszahlen sind normiert $0 \leq K \leq 1$. Je kleiner die Konditionszahl ist, eine desto schlechtere Kondition ist zu erwarten. Von einer Konditionszahl verlangt man, daß sie möglichst folgende fünf Eigenschaften erfüllt.

1. $0 \leq K \leq 1$
2. $K = 0$ genau dann, wenn $\det A = 0$.
3. $K = 1$ wird angenommen.
4. K ist invariant gegenüber Vertauschungen von Zeilen.
5. K ist invariant gegenüber Multiplikation von A mit einem Faktor ungleich Null.

Wir wollen nun eine Reihe von Konditionszahlen angeben.

I. Konditionszahl, die auf der Ungleichung von Hadamard beruht.

$$A = (a_{ij}) \quad ; \quad l_i = \left(\sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} \quad ; \quad V = \prod_{i=1}^{n} l_i$$

(V = Volumen eines Quaders mit den Seitenlängen l_i)

Man beachte : $V \geq |\det A|$; das „=“-Zeichen gilt genau dann, wenn die Zeilenvektoren paarweise aufeinander senkrecht stehen.

Definition :

$$K_H := \frac{|\det A|}{V}$$

Man sieht leicht, daß K_H die oben genannten fünf Eigenschaften erfüllt.

II. Konditionszahl der euklidschen Norm.

$$N(A) := (l_1^2 + \cdots + l_n^2)^{1/2} = \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{1/2}$$

$\overline{l} := \frac{N(A)}{\sqrt{n}}$; $\overline{l}$ ist das quadratische Mittel der Längen der Zeilenvektoren von A.

Definition :

$$K_N := \frac{|\det A|}{\overline{l}^n}$$

Für K_N gelten die Eigenschaften 1 bis 4 einer Konditionszahl.

Zusatz :

Berücksichtigt man $V \leq \overline{l}^n$, so erhält man : $K_N \leq K_H$.

III. Konditionszahl, die auf den singulären Werten von A beruht.

Definition :

$$K_O := \left(\frac{\Lambda_1}{\Lambda_n} \right)^{1/2}$$

Für K_O gelten die Eigenschaften 1,2,3,5 einer Konditionszahl.

Zusatz :

Es gilt die Beziehung : $K_O^n \leq K_N$

IV. Konditionszahl von Todd.

$\lambda_1, \lambda_2, \ldots, \lambda_n$ seien die Eigenwerte von A.

Definition :

$$K_{To} := \frac{\min_{i=1,\ldots,n} |\lambda_i|}{\max_{i=1,\ldots,n} |\lambda_i|}$$

Für K_{To} gelten die Eigenschaften 1,2,3 einer Konditionszahl.

V. Konditionszahl von Turing.

Definition :

$$K_{Tu} := \frac{n}{N(A)\, N(A^{-1})} = (K_N(A)\, K_N(A^{-1}))^{1/n}$$

Für K_{Tu} gelten die Eigenschaften 1,3,4 einer Konditionszahl.

VI. Definition :

$$K_i(A) := \begin{cases} \dfrac{1}{\|A\|_i \|A^{-1}\|_i} & \text{für } \det A \neq 0 \\ 0 & \text{für } \det A = 0 \end{cases}$$

Für $K_i(A)$ gelten alle Eigenschaften einer Konditionszahl, wenn man die Normen $i = 1,2,3$ zugrundelegt.

VII. Definition :

$$K_i^*(A) := \begin{cases} \dfrac{\alpha_i \max_{e \in \mathcal{E}} \|e\|_i}{\|A\|_i} & \text{für } \det A \neq 0 \\ 0 & \text{für } \det A = 0 \end{cases}$$

Für $K_i^*(A)$ gelten alle Eigenschaften einer Konditionszahl, wenn man die Normen $i = 1,2,3$ zugrundelegt.

III . K A P I T E L .

Numerische Auswertung der Ergebnisse

In den Ausführungen von Kapitel II haben wir keine Überlegungen angestellt über die numerische Auswertung der erzielten Ergebnisse. Es handelt sich bei gegebener Matrix A im wesentlichen um folgende Terme.

1. $\|A^{-1}\|_i \qquad i = 1,2,3$
2. $\max\limits_{i=1,\dots,n} \left(\sum\limits_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2} \qquad A^{-1} = (\alpha_{ij})$
3. $\max\limits_{e \in \mathcal{E}} \|A^{-1}e\|_1 \qquad i = 2,3$
4. $\sum\limits_{i=1}^{n} \left(\sum\limits_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$
5. $\left(\sum\limits_{i,j=1}^{n} \alpha_{ij}^2 \right)^{1/2}$

In den folgenden Ausführungen sollen Möglichkeiten angegeben werden, diese Ausdrücke numerisch auszuwerten. Wir gehen zunächst davon aus, daß die inverse Matrix A^{-1} explizit vorliegt (§ 1). Außerdem werden wir Verfahren angeben, welche ohne Kenntnis der inversen Matrix einen Teil dieser Terme zumindest näherungsweise zu berechnen gestatten (§ 2).

§ 1. Numerische Auswertung der Ergebnisse bei bekannter inverser Matrix A^{-1}.

Wir wollen in diesem Paragraphen davon ausgehen, daß die inverse Matrix $A^{-1} = (\alpha_{ij}) = (s_1\ s_2\ \dots\ s_n)$ explizit vorliegt. In diesem Fall bereiten nur die Terme $\|A^{-1}\|_3$ und $\max\limits_{e \in \mathcal{E}} \|A^{-1}e\|_i \quad i = 2,3$ Schwierigkeiten. Bei der Berechnung von $\|A^{-1}\|_3$ bedeutet die Kenntnis der inversen Matrix keinen wesentlichen Vorteil. Die Berechnung von $\|A^{-1}\|_3$ erfolgt im nächsten Paragraphen.

Berechnung von $\max\limits_{e \in \mathcal{E}} \|A^{-1}e\|_i \quad i = 2,3$.

Die Menge $\mathcal{E}$ besitzt 2^n Elemente. Man beachte, daß $\|A^{-1}e\|_i = \|A^{-1}(-e)\|_i$ gilt. Bei der exakten Berechnung von $\max\limits_{e \in \mathcal{E}} \|A^{-1}e\|_i$

hat man demnach 2^{n-1} Linearkombinationen der Spaltenvektoren von A^{-1} zu bilden und die Linearkombination mit maximaler Norm zu suchen. Für große n ist dies mit sehr hohem numerischen Aufwand verbunden. Es werden eine Reihe von Näherungsverfahren angegeben, welche die Berechnung von $\max_{e \in \mathcal{E}} \|A^{-1}e\|_i$ mit wesentlich geringerem numerischen Aufwand näherungsweise gestatten.

I. Auf Grund früherer Überlegung gilt :

$$\max_{e \in \mathcal{E}} \|A^{-1}e\|_i \leq \|A^{-1}\|_i \max_{e \in \mathcal{E}} \|e\|_i$$

Wir haben damit die Berechnung des Terms $\max_{e \in \mathcal{E}} \|A^{-1}e\|_i$ auf die Berechnung der entsprechenden Norm der inversen Matrix zurückgeführt.

II. Das folgende Näherungsverfahren beruht auf der Idee, einen Näherungsvektor e^* zu konstruieren, für den die Beziehung gilt:

$$\|A^{-1}e^*\|_i \approx \|A^{-1}e_0\|_i = \max_{e \in \mathcal{E}} \|A^{-1}e\|_i$$

Mit Hilfe der Spaltenvektoren $s_1, s_2, \ldots, s_n$ von A^{-1} konstruieren wir die Folge :

$$s^{(1)} := s_1 \quad ; \quad s^{(\nu+1)} = \begin{cases} s^{(\nu)} + s_{\nu+1} , & \text{falls } \|s^{(\nu)} + s_{\nu+1}\|_i \geq \|s^{(\nu)} - s_{\nu+1}\|_i \\ s^{(\nu)} - s_{\nu+1} , & \text{falls } \|s^{(\nu)} + s_{\nu+1}\|_i < \|s^{(\nu)} - s_{\nu+1}\|_i \end{cases}$$

Es gilt : $\|s^{(n)}\|_i \leq \max_{e \in \mathcal{E}} \|A^{-1}e\|_i \qquad \nu = 1, 2, \ldots, n$

Nach Konstruktion ist zu erwarten, daß $\|s^{(n)}\|_i$ im allgemeinen eine gute Näherung für $\max_{e \in \mathcal{E}} \|A^{-1}e\|_i$ darstellt.

III. Auch das folgende Näherungsverfahren beruht auf der Konstruktion eines Näherungsvektors e^*.

Wir teilen die n Spaltenvektoren $s_1, s_2, \ldots, s_n$ von A^{-1} in r etwa gleich große Abschnitte ein :

$$s_1, \ldots, s_n = s_1, \ldots, s_{n_1} ; s_{n_1+1}, \ldots, s_{n_1+n_2} ; \ldots ; s_{n_1+\ldots+n_{r-1}+1}, \ldots, s_{n_1+\ldots+n_r}$$

Dabei erhalten wir den Aufspaltungsvektor $u^T = (n_1, n_2, \ldots, n_r)$.

Für die einzelnen Abschnitte werden optimale Linearkombinationen gebildet.

$$\max_{e \in \mathcal{E}} \|(s_1 \ldots s_{n_1})e\|_i = \|(s_1 \ldots s_{n_1})e_1\|_i \quad ; \quad s^{(1)} := (s_1 \ldots s_{n_1})e_1$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\max_{e \in \mathcal{E}} \|(s_{n_1+\cdots+n_{r-1}} \ldots s_n)e\|_i = \|(s_{n_1+\cdots+n_{r-1}} \ldots s_n)e_r\|_i \quad ; \quad s^{(r)} := (s_{n_1+\cdots+n_{r-1}} \ldots s_n)e_r$$

e^* ist festgelegt durch die Forderung $\|A^{-1}e^*\|_i = \max_{e \in \mathcal{E}} \|(s^{(1)} \ldots s^{(r)})e\|_i$

Nach Konstruktion ist zu erwarten, daß $\|A^{-1}e^*\|_i$ eine gute Näherung für $\max_{e \in \mathcal{E}} \|A^{-1}e\|_i$ darstellt.

Für die Normen $i = 1,2,3$ lassen sich folgende im allgemeinen sehr groben Schranken angeben.

$$i = 1 : \quad \frac{1}{r} \max_{e \in \mathcal{E}} \|A^{-1}e\|_1 \le \|A^{-1}e^*\|_1 \le \max_{e \in \mathcal{E}} \|A^{-1}e\|_1$$

$$i = 2 : \quad \frac{1}{r} \max_{e \in \mathcal{E}} \|A^{-1}e\|_2 \le \|A^{-1}e^*\|_2 \le \max_{e \in \mathcal{E}} \|A^{-1}e\|_2$$

$$i = 3 : \quad \frac{1}{\sqrt{r}} \max_{e \in \mathcal{E}} \|A^{-1}e\|_3 \le \|A^{-1}e^*\|_3 \le \max_{e \in \mathcal{E}} \|A^{-1}e\|_3$$

Beweis :

Nach Konstruktion gilt für $i = 1,2$ die folgende Beziehung:

$$\max_{\nu=1,\ldots,r} \|s^{(\nu)}\|_{1,2} \le \|A^{-1}e^*\|_{1,2} \le \max_{e \in \mathcal{E}} \|A^{-1}e\|_{1,2} \le \|s^{(1)}\|_{1,2} + \cdots + \|s^{(r)}\|_{1,2} \le$$

$$r \max_{\nu=1,\ldots,r} \|s^{(\nu)}\|_{1,2}$$

Aus diesen Abschätzungen folgt unmittelbar die Behauptung :

Für $i = 3$ gilt:

$$\left(\sum_{\nu=1}^{r} \|s^{(\nu)}\|_3^2 \right)^{1/2} \le \|A^{-1}e^*\|_3 \le \max_{e \in \mathcal{E}} \|A^{-1}e\|_3 \le \|s^{(1)}\|_3 + \cdots + \|s^{(r)}\|_3$$

Nach Satz I.9 gilt für einen beliebigen Vektor x die Abschätzung $\sqrt{r}\,\|x\|_3 \ge \|x\|_2$. Daraus folgt unmittelbar die angegebene Abschätzung.

Zur Berechnung der Größen $\alpha_i = \dfrac{1}{\max\limits_{e \in \mathcal{E}} \|A^{-1}e\|_i}$, $\alpha_i^* = \dfrac{1}{\|A^{-1}e^*\|_i}$

und $\alpha_i^{**} = \dfrac{1}{\|A^{-1}\|_i \max\limits_{e \in \mathcal{E}} \|e\|_i}$ wurde ein Algolprogramm angefertigt.

Um uns einen Überblick über diese Größen zu verschaffen, geben wir einige Beispiele an.

1. Beispiel :

$$A = \begin{pmatrix} 10 & 0,1 & -0,1 & 0,1 & 0,1 & -0,1 & 0,0 & 0,0 \\ 0,0 & -10 & -0,1 & -0,1 & 0,1 & 0,0 & 0,1 & -0,1 \\ 0,1 & -0,1 & 11 & 0,0 & 0,0 & 0,1 & -0,1 & 0,1 \\ 0,0 & 0,0 & 0,1 & -12 & -0,1 & -0,1 & 0,0 & 0,1 \\ 0,0 & 0,1 & -0,1 & 0,0 & 13 & 0,1 & 0,1 & 0,0 \\ 0,0 & 0,1 & -0,1 & 0,1 & 0,1 & 14 & 0,1 & 0,0 \\ 0,1 & -0,1 & -0,1 & -0,1 & 0,1 & 0,1 & 15 & 0,1 \\ -0,1 & -0,1 & 0,1 & 0,1 & -0,1 & -0,1 & 0,0 & 16 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (3,3,2)$

	α_i	α_i^*	α_i^{**}
i = 1	$0,96\ 10^1$	$0,96\ 10^1$	$0,96\ 10^1$
i = 2	$0,15\ 10^1$	$0,15\ 10^1$	$0,12\ 10^1$
i = 3	$0,42\ 10^1$	$0,42\ 10^1$	$0,35\ 10^1$

Wir haben hier eine Matrix mit stark überwiegender Hauptdiagonale vorliegen. Die Größen α_i, α_i^* und α_i^{**} stimmen gut überein.

2. Beispiel :

$$A = \begin{pmatrix} 20 & -1,0 & 1,0 & 1,0 & 0,0 & -1,0 & 1,0 & 0,0 \\ 0,0 & -15 & 0,0 & -1,0 & 2,0 & 0,0 & 2,0 & 1,0 \\ 1,0 & -1,0 & 12 & 1,0 & 0,0 & -2,0 & 1,0 & -1,0 \\ 2,0 & -1,0 & -1,0 & -13 & 2,0 & 0,0 & 0,0 & 0,0 \\ 1,0 & -1,0 & -1,0 & 0,0 & 14 & 1,0 & -1,0 & 1,0 \\ -1,0 & 1,0 & 0,0 & 1,0 & 2,0 & 16 & 0,0 & 0,0 \\ 0,0 & 0,0 & 1,0 & -1,0 & -1,0 & -1,0 & 17 & 0,0 \\ 0,0 & 1,0 & 1,0 & 1,0 & 0,0 & 0,0 & -1,0 & 2,0 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (2,2,2,2)$

	α_i	α_i^*	α_i^{**}
i = 1	$0{,}82\ 10^1$	$0{,}82\ 10^1$	$0{,}82\ 10^1$
i = 2	$0{,}16\ 10^1$	$0{,}16\ 10^1$	$0{,}11\ 10^1$
i = 3	$0{,}45\ 10^1$	$0{,}45\ 10^1$	$0{,}38\ 10^1$

Wir haben hier eine Matrix mit stark überwiegender Hauptdiagonale vorliegen. Die Größen α_i, α_i^* und α_i^{**} stimmen gut überein.

3. Beispiel :

$$A = \begin{pmatrix} 10 & 1{,}0 & -1{,}0 & 2{,}0 & -2{,}0 & -1{,}0 & 1{,}0 & 0{,}0 \\ 10 & 1{,}0 & -1{,}0 & 2{,}0 & -2{,}0 & -1{,}0 & 1{,}0 & 0{,}01 \\ 0{,}1 & -0{,}1 & 10 & 0{,}1 & 0{,}0 & 0{,}0 & 0{,}1 & -0{,}1 \\ -0{,}1 & 0{,}0 & 0{,}0 & 11 & -0{,}1 & -0{,}1 & 0{,}1 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}1 & 12 & -0{,}1 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}1 & -0{,}1 & -0{,}1 & 0{,}1 & -12 & 0{,}1 & -0{,}1 \\ 0{,}1 & 0{,}0 & 0{,}0 & 0{,}1 & -0{,}1 & 0{,}0 & 13 & 0{,}0 \\ 0{,}0 & 0{,}0 & -0{,}1 & 0{,}1 & -0{,}1 & 0{,}0 & 0{,}0 & 14 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (3,3,2)$

	α_i	α_i^*	α_i^{**}
i = 1	$0{,}42\ 10^{-6}$	$0{,}42\ 10^{-6}$	$0{,}42\ 10^{-6}$
i = 2	$0{,}38\ 10^{-6}$	$0{,}38\ 10^{-6}$	$0{,}95\ 10^{-7}$
i = 3	$0{,}42\ 10^{-6}$	$0{,}42\ 10^{-6}$	$0{,}21\ 1o^{-6}$

Wir haben hier eine Matrix mit zwei fast linear abhängigen Zeilen verwendet. Die Größen α_i und α_i^{**} unterscheiden sich für i = 2,3 merklich voneinander.

4. Beispiel :

$$A = \begin{pmatrix} 10 & 10 & 0,0 & 0,0 & 0,1 & -0,1 & 0,1 & 0,1 \\ 2,0 & 2,0 & 0,0 & 0,1 & -0,1 & -0,1 & 0,1 & 0,1 \\ -2,0 & -2,0 & 10 & -0,1 & 0,1 & 0,1 & -0,1 & -0,1 \\ 0,1 & 0,1 & 0,1 & 11 & 0,0 & 0,1 & -0,1 & -0,1 \\ -0,1 & -0,1 & -0,1 & -0,1 & -12 & -0,1 & 0,1 & 0,0 \\ 0,0 & 0,01 & 0,1 & 0,1 & 0,1 & 13 & 0,1 & 0,0 \\ 0,0 & 0,0 & 0,1 & 0,1 & 0,1 & 0,1 & 14 & 0,0 \\ 0,0 & 0,0 & -0,1 & -0,1 & 0,0 & 0,0 & 0,0 & -14 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (3,3,2)$.

	α_i	α_i^*	α_i^{**}
i = 1	$0,50\ 10^{-4}$	$0,50\ 10^{-4}$	$0,50\ 10^{-4}$
i = 2	$0,25\ 10^{-4}$	$0,25\ 10^{-4}$	$0,39\ 10^{-5}$
i = 3	$0,35\ 10^{-4}$	$0,35\ 10^{-4}$	$0,15\ 10^{-4}$

Wir haben hier eine Matrix mit zwei fast linear abhängigen Spalten verwendet. Die Größen α_i und α_i^{**} unterscheiden sich für i = 2,3 merklich von einander.

5. Beispiel :

$$A = \begin{pmatrix} 1,0 & -2,0 & 3,0 & 0,0 & -5,0 & 10 & 7,0 & 1,0 \\ 2,0 & -3,0 & 0,0 & 1,0 & 5,0 & 9,0 & 2,0 & 0,0 \\ 0,0 & 0,0 & 1,0 & 2,0 & -3,0 & -1,0 & 2,0 & 5,0 \\ -1,0 & 3,0 & 2,0 & 1,0 & -1,0 & -3,0 & 2,0 & 1,0 \\ 5,0 & -6,0 & 0,0 & -6,0 & 7,0 & 1,0 & 8,0 & 3,0 \\ -1,0 & 2,0 & -3,0 & 1,0 & 2,0 & 7,0 & 0,0 & 1,0 \\ 8,0 & -1,0 & 2,0 & 3,0 & -3,0 & 0,0 & 5,0 & 1,0 \\ 0,0 & 1,0 & -1,0 & -2,0 & 3,0 & 8,0 & 5,0 & 9,0 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (2,2,2,2)$.

	α_i	α_i^*	α_i^{**}
i = 1	$0{,}88\ 10^0$	$0{,}10\ 10^1$	$0{,}88\ 10^0$
i = 2	$0{,}24\ 10^0$	$0{,}24\ 10^0$	$0{,}12\ 10^0$
i = 3	$0{,}57\ 10^0$	$0{,}57\ 10^0$	$0{,}46\ 10^0$

Wir haben die Koeffizienten der Matrix A beliebig eingegeben.
Die Größen α_i und α_i^{**} sind etwas unterschiedlich.

6. Beispiel :

$$A = \begin{pmatrix} 1{,}0 & 2{,}0 & -4{,}0 & -8{,}0 & 6{,}0 & 5{,}0 & 4{,}0 & -8{,}0 \\ 2{,}0 & -8{,}0 & 6{,}0 & -8{,}0 & -8{,}0 & 3{,}0 & 5{,}0 & 6{,}0 \\ 1{,}0 & -2{,}0 & 3{,}0 & 3{,}0 & 5{,}0 & -9{,}0 & 1{,}0 & -4{,}0 \\ -8{,}0 & 7{,}0 & -3{,}0 & -2{,}0 & 5{,}0 & 4{,}0 & 3{,}0 & 1{,}0 \\ -6{,}0 & 3{,}0 & 4{,}0 & 7{,}0 & 1{,}0 & -7{,}0 & -3{,}0 & -1{,}0 \\ 1{,}0 & 2{,}0 & 9{,}0 & 5{,}0 & -5{,}0 & -3{,}0 & -3{,}0 & 2{,}0 \\ 1{,}0 & 2{,}0 & -8{,}0 & 3{,}0 & 2{,}0 & -6{,}0 & 12 & 12 \\ -12 & 2{,}0 & 2{,}0 & -9{,}0 & 1{,}0 & 2{,}0 & -5{,}0 & 2{,}0 \end{pmatrix}$$

Als Aufspaltungsvektor verwenden wir $u^T = (3,3,2)$.

	α_i	α_i^*	α_i^{**}
i = 1	$0{,}85\ 10^0$	$0{,}86\ 10^0$	$0{,}85\ 10^0$
i = 2	$0{,}25\ 10^0$	$0{,}25\ 10^0$	$0{,}10\ 10^0$
i = 3	$0{,}60\ 10^0$	$0{,}60\ 10^0$	$0{,}50\ 10^0$

Wir haben die Koeffizienten der Matrix A beliebig eingegeben.
Die Größen α_i und α_i^{**} sind etwas von einander verschieden.

§ 2. Numerische Auswertung der Ergebnisse ohne Kenntnis der inversen Matrix A^{-1}.

Bei der Lösung sehr großer linearer Gleichungssysteme verwendet man im allgemeinen Iterationsverfahren. Die zusätzliche Berechnung der inversen Koeffizientenmatrix würde daher einen großen numerischen Mehraufwand erfordern. In den folgenden Ausführungen werden Möglichkeiten angegeben, einige der zu Beginn des Kapitels angegebenen Terme auch ohne Kenntnis der inversen Matrix zumindest näherungsweise zu berechnen.

Ausgangspunkt für weitere Überlegungen bildet der folgende Satz.

Satz III.1 :

Es seien $\| \ \|_i$ und $\| \ \|_j$ zwei beliebige Vektornormen; A $\det A \neq 0$ eine beliebige Matrix, dann gilt die folgende Beziehung :

$$\max_{\|r\|_j = 1} \|A^{-1} r\|_i = \frac{1}{\min_{\|s\|_i = 1} \|As\|_j} \quad ; \quad A^{-1} = (\alpha_{ij})$$

Beweis :

I. Es existiert ein Vektor r_o mit $\|r_o\|_j = 1$ und $\max_{\|r\|_j = 1} \|A^{-1} r\|_i = \|A^{-1} r_o\|_i$

Wir setzen $s_1 := \dfrac{A^{-1} r_o}{\|A^{-1} r_o\|_i}$. Es gelten die Gleichungen :

$$\|s_1\|_i = 1 \quad \text{und} \quad \|As_1\|_j = \frac{\|r_o\|_j}{\|A^{-1} r_o\|_i} = \frac{1}{\max_{\|r\|_j = 1} \|A^{-1} r\|_i}$$

Wegen $\|As_1\|_j \geq \min_{\|s\|_i = 1} \|As\|_j$ gilt : $\max_{\|r\|_j = 1} \|A^{-1} r\|_i \leq \dfrac{1}{\min_{\|s\|_i = 1} \|As\|_j}$

II. Es existiert ein Vektor s_o mit $\|s_o\|_i = 1$ und $\min_{\|s\|_i = 1} \|As\|_j = \|As_o\|_j$

Wir setzen $r_1 := \dfrac{As_o}{\|As_o\|_j}$. Es gelten die Gleichungen :

$$\|r_1\|_j = 1 \quad \text{und} \quad \|A^{-1} r_1\|_i = \frac{\|s_o\|_i}{\|As_o\|_j} = \frac{1}{\min_{\|s\|_i = 1} \|As\|_j} \ .$$

Wegen $\|A^{-1}r_1\|_i \leq \max_{\|r\|_j=1} \|A^{-1}r\|_i$ gilt : $\max_{\|r\|_j=1} \|A^{-1}r\|_i \geq \frac{1}{\min_{\|s\|_i=1} \|As\|_j}$

Aus I. und II. folgt die Behauptung des Satzes.

In der folgenden Zusammenstellung wird das Ergebnis des Satzes für die Normen $i = 1,2,3$ explizit angegeben.

$$\min_{\|s\|_1=1} \|As\|_1 = \frac{1}{\|A^{-1}\|_1} \; ; \; \min_{\|s\|_1=1} \|A^T s\|_1 = \frac{1}{\|A^{-1}\|_2} \; ; \; \min_{\|s\|_2=1} \|As\|_1 = \alpha_2$$

$$\min_{\|s\|_3=1} \|As\|_1 = \alpha_3 \; ; \; \min_{\|s\|_1=1} \|As\|_2 = \frac{1}{\max_{j=1,\dots,n} |\alpha_{ij}|} \; ; \; \min_{\|s\|_1=1} \|A^T s\|_2 = \frac{1}{\max_{i=1,\dots,n} |\alpha_{ij}|}$$

$$\min_{\|s\|_2=1} \|As\|_2 = \frac{1}{\|A^{-1}\|_2} \; ; \; \min_{\|s\|_2=1} \|A^T s\|_2 = \frac{1}{\|A^{-1}\|_1} \; ; \; \min_{\|s\|_3=1} \|As\|_2 = \frac{1}{\max_{j=1,\dots,n} \left(\sum_{i=1}^{n} \alpha_{ij}^2 \right)^{1/2}}$$

$$\min_{\|s\|_3=1} \|A^T s\|_2 = \frac{1}{\max_{i=1,\dots,n} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}} \; ; \; \min_{\|s\|_1=1} \|As\|_3 = \frac{1}{\max_{i=1,\dots,n} \left(\sum_{j=1}^{n} \alpha_{ij}^2 \right)^{1/2}}$$

$$\min_{\|s\|_1=1} \|A^T s\|_3 = \frac{1}{\max_{j=1,\dots,n} \left(\sum_{i=1}^{n} \alpha_{ij}^2 \right)^{1/2}} \; ; \; \min_{\|s\|_2=1} \|A^T s\|_3 = \alpha_3 \; ; \; \min_{\|s\|_3=1} \|As\|_3 = \frac{1}{\|A^{-1}\|_3}$$

Beweis :

Die meisten der angegebenen Gleichungen lassen sich unmittelbar aus dem letzten Satz folgern. Wir wollen uns hier auf den Beweis der einzigen nichttrivialen Gleichung beschränken: $\min_{\|s\|_2=1} \|A^T s\|_3 = \alpha_3$

$$\text{Zu zeigen : } \max_{\|s\|_3=1} \|A^{-1^T} s\|_2 = \max_{e \in \mathcal{E}} \|A^{-1} e\|_3$$

$s_1, s_2, \dots, s_n$ seien die Spaltenvektoren von A^{-1}. Es existieren Vektoren $e_o \in \mathcal{E}$ und s^* mit $\|s^*\|_3 = 1$, so daß gilt :

$$\max_{\|s\|_3=1} \|A^{-1^T} s\|_2 = \|A^{-1^T} s^*\|_2 \; ; \; \max_{e \in \mathcal{E}} \|A^{-1} e\|_3 = \|A^{-1} e_o\|_3$$

Es gelten die beiden folgenden Darstellungen :

$$\|A^{-1^T} s^*\|_2 = |s_1^T s^*| + \dots + |s_n^T s^*| = (\varepsilon_1 s_1^T + \dots + \varepsilon_n s_n^T) s^* \; ; \; \varepsilon_i = \operatorname{sgn} s_i^T s^*$$

$$\|A^{-1} e_o\|_3 = \|\varepsilon_1^{(o)} s_1 + \dots + \varepsilon_n^{(o)} s_n\|_3 = (\varepsilon_1^{(o)} s_1 + \dots + \varepsilon_n^{(o)} s_n) s_o \text{ , wobei}$$

$$e_0 = (\varepsilon_i^{(0)}) \quad \text{und} \quad s_0 = \frac{\sum_{\nu=1}^{n} \varepsilon_\nu^{(0)} s_\nu}{\| \sum_{\nu=1}^{n} \varepsilon_\nu^{(0)} s_\nu \|_3}$$

Es gelten die beiden folgenden Ungleichungen :

$$\| A^{-1} s^T s^* \|_2 = \Big(\sum_{\nu=1}^{n} \varepsilon_\nu s_\nu^T \Big) s^* \le \| \sum_{\nu=1}^{n} \varepsilon_\nu s_\nu \|_3 \le \max_{e \in \mathfrak{E}} \| A^{-1} e \|_3$$

$$\| A^{-1} s^T s^* \|_2 \ge \| A^{-1} s_0^T \|_2 = \sum_{\nu=1}^{n} | s_\nu^T s_0 | \ge \sum_{\nu=1}^{n} \varepsilon_\nu^{(\nu)} s_\nu^T s_0 = \| A^{-1} e_0 \|_3$$

Aus beiden Ungleichungen folgt die Behauptung.

Der letzte Satz gibt uns die Möglichkeit, Terme, welche die inverse Matrix explizit enthalten, nur mit Hilfe der bekannten Ausgangsmatrix darzustellen.

Die folgenden Überlegungen gelten der Konstruktion von Iterationsverfahren, mit deren Hilfe man Ausdrücke der Art $\min_{\|s\|_j = 1} \|As\|_i$ und $\max_{\|s\|_j = 1} \|As\|_i$ numerisch auswerten kann.

Der Grundgedanke für die folgenden Iterationsverfahren besteht darin, daß man eine geeignete Folge von Vektoren $(s_\kappa)_{\kappa = 0,1,\ldots}$ mit $\|s_\kappa\|_j = 1$ konstruiert, für die bei der Berechnung von $\min_{\|s\|_j = 1} \|As\|_i$ gilt :

$$\|As_{\kappa+1}\|_i < \|As_\kappa\|_i$$

bzw. bei der Berechnung von $\max_{\|s\|_j = 1} \|As\|_i$ gilt :

$$\|As_{\kappa+1}\|_i > \|As_\kappa\|_i$$

Es ist nicht gelungen, für alle Vektornormen $\| \ \|_i$ und $\| \ \|_j$ brauchbare Iterationsverfahren zu entwickeln. Wir werden jetzt einige Beispiele angeben, bei denen gute Ergebnisse erzielt werden können.

I. Iterative Berechnung von $\min_{\|s\|_3 = 1} \|As\|_3 = \frac{1}{\|A^{-1}\|_3}$

Die Aufgabe besteht in der Konstruktion einer geeigneten Vektorfolge $(s_\kappa)_{\kappa = 0,1,\ldots}$.

Wir machen den Ansatz : $(*)\ s_{\kappa+1} = \frac{s_\kappa + \lambda_\kappa v_\kappa}{\| s_\kappa + \lambda_\kappa v_\kappa \|_3}$

Die Iteration ist durch folgende Angaben festgelegt:

1. Angabe eines Startvektors s_0 mit $\|s_0\|_3 = 1$
2. Angabe eines Korrekturvektors v_κ und eines Skalarfaktors λ_κ bei jedem Iterationsschritt.

Zu 1 : Es ist bequem, mit einem Einheitsvektor zu starten. Wir setzen $v_0 := e_{\nu_0}$ mit $\|Ae_{\nu_0}\|_3 = \min_{\nu=1,\dots,n} \|Ae_\nu\|_3$.

Zu 2 : Bei der Wahl der Größen λ_κ und v_κ beachten wir die Forderung :

$$\|As_{\kappa+1}\|_3 = \frac{\|As_\kappa + \lambda_\kappa Av_\kappa\|_3}{\|s_\kappa + \lambda_\kappa v_\kappa\|_3} < \|As_\kappa\|_3$$

Wir führen die folgenden Abkürzungen ein :

$$u_\kappa := As_\kappa \quad ; \quad \omega_\kappa := \sqrt{u_\kappa^T u_\kappa}$$

Für die weiteren Betrachtungen ist es günstig, von der folgenden äquivalenten Forderung auszugehen :

$$M(\lambda_\kappa, v_\kappa) := \frac{(As_\kappa + \lambda_\kappa Av_\kappa)^2}{(s_\kappa + \lambda_\kappa v_\kappa)^2} < (As_\kappa)^2 = \omega_\kappa^2$$

Durch Ausmultiplizieren erhält man die folgende Gleichung :

$$M(\lambda_\kappa, v_\kappa) = \frac{\omega_\kappa^2 + 2\lambda_\kappa u_\kappa^T Av_\kappa + \lambda_\kappa^2 v_\kappa^T A^T Av_\kappa}{1 + 2\lambda_\kappa s_\kappa^T v_\kappa + \lambda_\kappa^2 v_\kappa^T v_\kappa}$$

Der Differentialquotient nach λ_κ im Nullpunkt ergibt :

$$M'_{|\lambda_\kappa}(0, v_\kappa) = 2v_\kappa^T (A^T u_\kappa - \omega_\kappa^2 s_\kappa)$$

Für v_κ mit $\|v_\kappa\|_3 = 1$ ist $M'_{|\lambda_\kappa}(0, v_\kappa)$ ein Maß für die Empfindlichkeit der Größe ω_κ gegenüber Änderungen in Richtung des Vektors v_κ .

<u>Folgerung :</u>

1. Die optimale Wahl für den Korrekturvektor v_κ ist bis auf einen Faktor eindeutig bestimmt.

   Wir wählen $v_\kappa := A^T u_\kappa - \omega_\kappa^2 s_\kappa$

2. Die Forderung $\omega_{\kappa+1} < \omega_\kappa$ läßt sich durch den Ansatz $(*)$ genau dann nicht erfüllen, wenn für alle Vektoren v_κ mit $\|v_\kappa\|_3 = 1$ gilt : $M^1_{\lambda_\kappa}(0, v_\kappa) = 0$. Dies ist genau dann der Fall, wenn gilt :

$A^T u_\kappa - \omega_\kappa^2 s_\kappa = (A^T A - \omega_\kappa^2 E) s_\kappa = 0$. Dies bedeutet : s_κ ist Eigenvektor und ω_κ^2 ist Eigenwert von $A^T A$.

Die folgenden Überlegungen gelten einer geeigneten Wahl des Skalarfaktors λ_κ.

Hilfssatz III.1 :

Für $v_\kappa = A^T u_\kappa - \omega_\kappa^2 s_\kappa$ gelten die folgenden Relationen :

I. $v_\kappa^T s_\kappa = 0$; II. $Av_\kappa = (AA^T - \omega_\kappa^2 E) u_\kappa$; III. $u_\kappa^T A v_\kappa = v_\kappa^T v_\kappa$

Beweis :

Zu I : $v_\kappa^T s_\kappa = s_\kappa^T (A^T u_\kappa - \omega_\kappa^2 s_\kappa) = s_\kappa^T A^T u_\kappa - \omega_\kappa^2 = u_\kappa^T u_\kappa - \omega_\kappa^2 = 0$

Zu II : $Av_\kappa = AA^T u_\kappa - \omega_\kappa^2 A s_\kappa = AA^T u_\kappa - \omega_\kappa^2 u_\kappa = (AA^T - \omega_\kappa^2 E) u_\kappa$

zu III. : $v_\kappa^T A^T u_\kappa = u_\kappa^T A A^T u_\kappa - \omega_\kappa^2 s_\kappa^T A^T u_\kappa = u_\kappa^T A A^T u_\kappa - \omega_\kappa^4$

$v_\kappa^T v_\kappa = u_\kappa^T A A^T u_\kappa - 2\omega_\kappa^2 s_\kappa^T A^T u_\kappa + \omega_\kappa^4 = u_\kappa^T A A^T u_\kappa - \omega_\kappa^4$

Bei geeigneter Wahl des Skalarfaktors λ_κ erhält man monotone Konvergenz der Folge $(\omega_\kappa)_{\kappa = 0,1,\ldots}$. Eine hinreichende Bedingung hierfür liefert der folgende Satz :

Satz III.2 :

Es sei $v_\kappa = A^T u_\kappa - \omega_\kappa^2 s_\kappa \neq \sigma$; $\lambda_\kappa < 0$

$$|\lambda_\kappa| < \begin{cases} \dfrac{2}{\dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2} & , \text{ falls } \dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2 > 0 \\ \infty & , \text{ falls } \dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2 \leq 0 \end{cases}$$

Dann gilt : $\omega_{\kappa+1}^2 = M(\lambda_\kappa, v_\kappa) < \omega_\kappa^2$

Beweis :

Berücksichtigt man Hilfssatz III.1 , so erhält man :

$$M(\lambda_\kappa, v_\kappa) = \frac{\omega_\kappa^2 + 2\lambda_\kappa v_\kappa^T v_\kappa + \lambda_\kappa^2 v_\kappa^T A^T A v_\kappa}{1 + \lambda_\kappa^2 v_\kappa^T v_\kappa}$$

Daraus ergeben sich die folgenden Relationen :

$$M(\lambda_\kappa, v_\kappa) < \omega_\kappa^2 \iff 2\lambda_\kappa v_\kappa^T v_\kappa + \lambda_\kappa^2 v_\kappa^T A^T A v_\kappa < \lambda_\kappa^2 v_\kappa^T v_\kappa \cdot \omega_\kappa^2 \iff$$

$$\lambda_\kappa^2 \left[\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2 \right] \leq -2\lambda_\kappa$$

Wegen $\lambda_\kappa < 0$ erhält man hieraus die folgende Beziehung :

$$|\lambda_\kappa| \left[\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2 \right] \leq 2$$

Diese Relation ergibt unmittelbar die Behauptung des Satzes.

Zusatz :

Für λ_κ lassen sich die beiden folgenden Schranken angeben :

$$|\lambda_\kappa| < \frac{2}{\Lambda_n - \omega_\kappa^2} \quad \text{bzw.} \quad |\lambda_\kappa| < \frac{2}{\Lambda_n - \Lambda_1}$$

Beweis :

Man beachte : 1. Nach Satz I.1 gilt : $\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} \leq \Lambda_n$

2. Nach Konstruktion von ω_κ gilt : $\Lambda_1 \leq \omega_\kappa^2 \leq \Lambda_n$

Der letzte Satz hat eine hinreichende Bedingung für die monotone Konvergenz der Folge $(\omega_\kappa)_{\kappa=0,1,\ldots}$ bezüglich der Wahl des Skalarfaktors λ_κ geliefert. Die optimale Wahl für λ_κ ist bestimmt durch die Forderung $M(\lambda_\kappa, v_\kappa) \overset{!}{=} \min$. Durch Auswertung der Gleichung $M'_{\lambda_\kappa}(\lambda_\kappa, v) = 0$ erhält man für den optimalen Skalarfaktor λ_κ :

$$\lambda_\kappa = \frac{\left[\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right] - \sqrt{\left[\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right] + 4 v_\kappa^T v_\kappa}}{2 v_\kappa^T v_\kappa}$$

Bei der Ausrechnung beachte man die Ergebnisse von Hilfssatz III.1

Die Iteration zur Berechnung von $\min_{\|s\|_3=1} \|As\|_3$ ist nun eindeutig festgelegt. Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.3 :

A sei eine reelle reguläre Matrix. $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ mit $x^{(i)T} x^{(j)} = \delta_{ij}$ sei ein orthonormiertes System von Eigenvektoren der symmetrischen positiv definiten Matrix $A^T A$ und $0 < \mu_1 < \mu_2 < \ldots < \mu_n$ seien die zugehörigen Eigenwerte. Der Startvektor s_0 der Iteration besitze die folgende Darstellung :

$$s_0 = \sum_{\nu=1}^{n} c_\nu^{(0)} x^{(\nu)} \quad \text{mit} \quad c_1^{(0)} \neq 0$$

Für den Skalarfaktor λ_κ der Iteration gelte die Beziehung :

$$\lambda_\kappa < 0 \; ; \; 0 < \varepsilon \leq |\lambda_\kappa| \leq \frac{2}{\left|\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right|} - \varepsilon \quad \text{mit } \varepsilon \leq \frac{1}{\mu_n}$$

Dann konvergiert die Folge $\omega_\kappa \xrightarrow[\kappa \to \infty]{} \sqrt{\mu_1} = \min_{\|s\|_3=1} \|As\|_3$

Beweis :

Die Folge $(\omega_\kappa)_{\kappa=0,1,\ldots}$ ist konvergent, da sie monoton und beschränkt ist. Wegen $0 < \varepsilon \leq |\lambda_\kappa|$ und nach Konstruktion der Iterationsfolge gilt : $M'_{|\lambda_\kappa}(0, v_\kappa) \xrightarrow[\kappa \to \infty]{} 0$. Daraus folgt : $A^T A s_\kappa - \omega_\kappa^2 s_\kappa \xrightarrow[\kappa \to \infty]{} 0$

Daraus folgt : 1. $s_\kappa \xrightarrow[\kappa \to \infty]{} \pm x^{(\nu)}$, 2. $\omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} \mu_\nu$

Die Behauptung des Satzes lautet : $\mu_\nu = \mu_1$

Wir zeigen dies durch einen indirekten Beweis.

Annahme : $\mu_\nu > \mu_1$

Für die Vektoren s_κ existiert die folgende Darstellung :

$$(*) \qquad s_\kappa = \sum_{\nu=1}^{n} c_\nu^{(\kappa)} x^{(\nu)}$$

Man beachte, daß aus $\mu_\nu > \mu_1$ folgt : $c_1^{(\kappa)} \xrightarrow[\kappa \to \infty]{} 0$

Die Aussage $c_1^{(\kappa)} \xrightarrow[\kappa \to \infty]{} 0$ führen wir zu einem Widerspruch.

$$s_{\kappa+1} = \frac{s_\kappa + \lambda_\kappa v_\kappa}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$$

Wegen $v_\kappa = A^T A s_\kappa - \omega_\kappa^2 s_\kappa$ erhält man die folgende Beziehung :

$$s_{\kappa+1} = \frac{s_\kappa + \lambda_\kappa (A^T A - \omega_\kappa^2 E) s_\kappa}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$$

Durch Einsetzen von Darstellung $(*)$ erhält man die Gleichung :

$$\sum_{\nu=1}^{n} c_\nu^{(\kappa+1)} x^{(\nu)} = \frac{\sum_{\nu=1}^{n} c_\nu^{(\kappa)} x^{(\nu)} + \lambda_\kappa (A^T A - \omega_\kappa^2 E) \sum_{\nu=1}^{n} c_\nu^{(\kappa)} x^{(\nu)}}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$$

Berücksichtigt man die Relation $\sum_{\nu=1}^{n} c_\nu^{(\kappa)} A^T A x^{(\nu)} = \sum_{\nu=1}^{n} c_\nu^{(\kappa)} \mu_\nu x^{(\nu)}$, so erhält man durch Koeffizientenvergleich die folgende Beziehung :

$$c_\nu^{(\kappa+1)} = \frac{c_\nu^{(\kappa)} + \lambda_\kappa (\mu_\nu - \omega_\kappa^2) c_\nu^{(\kappa)}}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$$

Insbesondere gilt : $\dfrac{c_1^{(\kappa+1)}}{c_1^{(\kappa)}} = \dfrac{1 + \lambda_\kappa (\mu_1 - \omega_\kappa^2)}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$

Wegen $\lambda_\kappa < 0$ und $\mu_1 - \omega_\kappa^2 < 0$ gilt : $c_1^{(\kappa)} \neq 0$ für $\kappa = 0,1,\ldots$, außerdem ändert $c_1^{(\kappa)}$ sein Vorzeichen nicht.

Wir zeigen nun $\dfrac{c_1^{(\kappa+1)}}{c_1^{(\kappa)}} > 1$ für $\kappa \geq K$

Dies ist ein Widerspruch zu $c_1^{(\kappa)} \xrightarrow[\kappa \to \infty]{} 0$

Es gelten - wie unmittelbar ersichtlich - die folgenden Relationen :

$$\frac{c_1^{(\kappa+1)}}{c_1^{(\kappa)}} > 1 \iff 1 + \lambda_\kappa^2 v_\kappa^T v_\kappa < (1 + \lambda_\kappa(\mu_1 - \omega_\kappa^2))^2 \iff$$

$$\lambda_\kappa^2 (v_\kappa^T v_\kappa - (\mu_1 - \omega_\kappa^2)^2) < 2\lambda_\kappa (\mu_1 - \omega_\kappa^2)$$

Wegen $\lambda_\kappa < 0$ und $\mu_1 - \omega_\kappa^2 < 0$ ist dies äquivalent zu folgender Relation :

$$|\lambda_\kappa| (v_\kappa^T v_\kappa - (\mu_1 - \omega_\kappa^2)^2) < 2 |\mu_1 - \omega_\kappa^2|$$

Wegen $M'_{|\lambda_\kappa}(0, v_\kappa) = 2 v_\kappa^T v_\kappa$ gilt : $v_\kappa^T v_\kappa \xrightarrow[\kappa \to \infty]{} 0$

Beachtet man $|\mu_1 - \omega_\kappa^2| \geq |\mu_1 - \mu_\nu| > 0$, so ist die obige Ungleichung für $\kappa \geq K$ erfüllt, wobei K geeignet zu wählen ist. Dies ist ein Widerspruch zu $c_1^{(\kappa)} \xrightarrow[\kappa \to \infty]{} 0$ und damit zu $\mu_\nu > \mu_1$.

Bemerkung :

Das eben angegebene Iterationsverfahren steht in gewisser Analogie zu einem bekannten Gradientenverfahren (vgl. Faddajew [3]) zur Berechnung des größten bzw. kleinsten Eigenwertes einer positiv definiten symmetrischen Matrix. Die Anwendung dieses Gradientenverfahrens würde die Berechnung von $A^T A$ erfordern. Das angegebene Iterationsverfahren kommt ohne die für große Matrizen aufwendige Berechnung von $A^T A$ aus und ist daher vorzuziehen.

Numerischer Aufwand des Verfahrens :

Bei jedem Iterationsschritt ist bei der Berechnung des Korrekturvektors v_κ im wesentlichen der Term $A^T u_\kappa$ und bei der Berechnung des optimalen Skalarfaktors λ_κ im wesentlichen der Term Av_κ zu berechnen. Die Anzahl der Iterationen hängt ab von der Güte des Startvektors und der Struktur der Matrix A. Das angegebene Verfahren ist günstig im Vergleich zu bekannten Verfahren.

Wir werden jetzt ein weiteres Iterationsverfahren zur Berechnung von $\min_{\|Su\|_3 = 1} \|Au\|_3$ angeben, welches mit wesentlich geringerem numerischen Aufwand verbunden ist. Allerdings wird der Grenzwert der Iteration im allgemeinen nur ein Näherungswert des gesuchten Wertes sein.

Wir gehen wieder aus von dem folgenden Ansatz :

$$(*) \qquad s_{\kappa+1} = \frac{s_\kappa + \lambda_\kappa v_\kappa}{\| s_\kappa + \lambda_\kappa v_\kappa \|_3} \quad ; \quad s_\kappa = (s_i^{(\kappa)})$$

Wir starten wieder mit einem Einheitsvektor :

$$v_0 := e_{\nu_0} \quad \text{mit} \quad \| Ae_{\nu_0} \|_3 = \min_{\nu=1,\dots,n} \| Ae_\nu \|_3$$

Für den Korrekturvektor v_κ nehmen wir ebenfalls einen Einheitsvektor :

$$v_\kappa := e_{\mu_\kappa}$$

$$\mu_0 = \begin{cases} \nu_0 + 1 & , \text{ falls } \nu_0 < n \\ 1 & , \text{ falls } \nu_0 = n \end{cases} \quad ; \quad \mu_{\kappa+1} = \begin{cases} \mu_\kappa + 1 & , \text{ falls } \mu_\kappa < n \\ 1 & , \text{ falls } \mu_\kappa = n \end{cases}$$

Dies bedeutet, daß wir bei jedem Iterationsschritt genau eine Komponente des Vektors s_κ ändern und anschließend normieren. Die Berechnung des Korrekturvektors v_κ ist also mit sehr geringem numerischen Aufwand verbunden.

Wir führen wieder die folgenden Abkürzungen ein :

$$A = (a_1 \; a_2 \; \dots \; a_n) \quad ; \quad u_\kappa := As_\kappa \; ; \; \omega_\kappa = \| u_\kappa \|_3$$

Die Wahl des Skalarfaktors λ_κ wird durch die folgende Forderung festgelegt :

$$\frac{\| A(s_\kappa + \lambda_\kappa e_{\mu_\kappa}) \|_3}{\| s_\kappa + \lambda_\kappa e_{\mu_\kappa} \|_3} \overset{!}{=} \min \iff$$

$$M_{\mu_\kappa}(\lambda_\kappa) := \frac{(As_\kappa + \lambda_\kappa a_{\mu_\kappa})^2}{(s_\kappa + \lambda_\kappa e_{\mu_\kappa})^2} = \frac{\omega_\kappa^2 + 2\lambda_\kappa u_\kappa^T a_{\mu_\kappa} + \lambda_\kappa a_{\mu_\kappa}^T a_{\mu_\kappa}}{1 + 2\lambda_\kappa s_{\mu_\kappa}^{(\kappa)} + \lambda_\kappa^2} \overset{!}{=} \min$$

Nach elementarer Rechnung erhält man die Beziehung :

$$M'_{\mu_\kappa}(0) = 2(u_\kappa^T a_{\mu_\kappa} - s_{\mu_\kappa}^{(\kappa)} \omega_\kappa^2)$$

Die Forderung $\omega_{\kappa+1} < \omega_\kappa$ läßt sich genau dann erfüllen, wenn $M'_{\mu_\kappa}(0) \neq 0$.

Durch Auswertung der Gleichung $M'_{\mu_\kappa}(\lambda_\kappa) = 0$ erhält man den optimalen Skalarfaktor λ_κ.

Wir wählen die folgenden Abkürzungen :

$$\gamma_\nu := a_\nu^T a_\nu \; ; \quad S_\kappa := a_{\mu_\kappa}^T u_\kappa - s_{\mu_\kappa}^{(\kappa)} \omega_\kappa^2 \; ; \quad L_\kappa := a_{\mu_\kappa}^T u_\kappa - s_{\mu_\kappa}^{(\kappa)} \gamma_{\mu_\kappa}$$

$$\lambda_\kappa = \begin{cases} \frac{1}{2}\left[\left(\frac{\gamma_{\mu_\kappa} - \omega_\kappa}{L_\kappa}\right) - \operatorname{sgn} L_\kappa \sqrt{\left(\frac{\gamma_{\mu_\kappa} - \omega_\kappa}{L_\kappa}\right)^2 + 4\,\frac{S_\kappa}{L_\kappa}}\right], & \text{falls } L_\kappa \neq 0 \\ -\frac{S_\kappa}{\gamma_{\mu_\kappa} - \omega_\kappa}, & \text{falls } L_\kappa = 0 \end{cases}$$

Nach diesen Ausführungen ist das Iterationsverfahren eindeutig festgelegt.

Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.4 :

A sei eine reelle reguläre Matrix. $0 < \mu_1 < \mu_2 < \dots < \mu_n$ seien die Eigenwerte von A^TA. Dann konvergiert die Folge $\omega_\kappa \xrightarrow[\kappa \to \infty]{} \sqrt{\mu_i}$

Beweis :

Nach Konstruktion der Iteration gilt : $M'_{\mu_\kappa}(o) \xrightarrow[\kappa \to \infty]{} 0$.

Daraus folgt : $a_{\mu_\kappa}^T u_\kappa - s_{\mu_\kappa}^{(\kappa)} \omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} 0$ für $\mu_\kappa = 1,2,\dots,n$.

Daraus folgt : $A^T u_\kappa - \omega_\kappa^2 s_\kappa \xrightarrow[\kappa \to \infty]{} 0$ Daraus folgt :

$(AA^T - \omega_\kappa^2 E)s_\kappa \xrightarrow[\kappa \to \infty]{} 0$. Daraus folgt : $s_\kappa \xrightarrow[\kappa \to \infty]{} \pm x^{(i)}$

$\omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} \mu_i$

Bemerkung :

$(\omega_\kappa)_{\kappa = 0,1,\dots}$ konvergiert also nicht notwendig gegen die Wurzel des kleinsten Eigenwertes von A^TA . Experimentelle Untersuchungen haben ergeben, daß die Folge im allgemeinen gegen die Wurzel eines kleinen Eigenwertes von A^TA konvergiert.

Die Bedeutung des eben angegebenen Iterationsverfahrens liegt in der Möglichkeit einer geeigneten Kombination beider Iterationsverfahren, wodurch der numerische Aufwand des ersten Verfahrens wesentlich verringert werden kann.

Numerischer Aufwand der Iteration :

Nach einmaliger Berechnung der Werte $\gamma_\nu = a_\nu^T a_\nu \quad \nu = 1,2,\ldots,n$ sind bei jedem Iterationsschritt im wesentlichen die Skalarprodukte $a_{\mu_\kappa}^T u_\kappa$ und $u_\kappa^T u_\kappa$ zu bilden. Im allgemeinen kommt man mit einigen n Iterationen aus. Das Iterationsverfahren hat also den Vorteil eines geringen numerischen Aufwandes.

II. Iterative Berechnung von $\max_{\|s\|_3=1} \|As\|_3 = \|A\|_3$.

Die hier angegebenen Iterationsverfahren verlaufen in völliger Analogie zu den eben diskutierten. Wir behalten die bisherige Bezeichnungsweise bei und begnügen uns mit einer knappen Darstellung der Iterationsverfahren.

Ansatz : $(*) \qquad s_{\kappa+1} = \dfrac{s_\kappa + \lambda_\kappa v_\kappa}{\|s_\kappa + \lambda_\kappa v_\kappa\|_3}$

Wir starten wieder mit einem Einheitsvektor :

$$v_0 := e_{\nu_0} \text{ mit } \|Ae_{\nu_0}\|_3 = \max_{\nu=1,\ldots,n} \|Ae_\nu\|_3$$

Bei der Wahl der Größen λ_κ und v_κ beachten wir die Forderung :

$$\|As_{\kappa+1}\|_3 = \frac{\|As_\kappa + \lambda_\kappa Av_\kappa\|_3}{\|s_\kappa + \lambda_\kappa v_\kappa\|_3} > \|As_\kappa\|_3$$

Wir betrachten wieder die äquivalente Forderung :

$$M(\lambda_\kappa, v_\kappa) := \frac{(As_\kappa + \lambda_\kappa Av_\kappa)^2}{(s_\kappa + \lambda_\kappa v_\kappa)^2} = \omega_{\kappa+1}^2 > (As_\kappa)^2 = \omega_\kappa^2$$

Der Differentialquotient nach λ_κ im Nullpunkt ergibt :

$$M'_{\lambda_\kappa}(0, v_\kappa) = 2v_\kappa^T (Au_\kappa - \omega_\kappa^2 s_\kappa)$$

Wir wählen $v_\kappa := \tilde{A}u_\kappa - \omega_\kappa^2 s_\kappa$

Bei geeigneter Wahl des Skalarfaktors λ_κ erhält man monotone Konvergenz der Folge $(\omega_\kappa)_{\kappa=0,1,\ldots}$. Eine hinreichende Bedingung hierfür liefert der folgende Satz :

Satz III.5 :

Es sei $v_\kappa = \tilde{A}u_\kappa - \omega_\kappa^2 s_\kappa \neq \theta$; $\lambda_\kappa > 0$

$$\lambda_\kappa < \begin{cases} \dfrac{2}{\omega_\kappa^2 - \dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa}} & , \text{falls } \omega_\kappa^2 - \dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} > 0 \\ \infty & , \text{falls } \omega_\kappa^2 - \dfrac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} \leq 0 \end{cases}$$

Dann gilt: $M(\lambda_\kappa, v_\kappa) = \omega_{\kappa+1}^2 > \omega_\kappa^2$

Beweis :

Durch elementare Rechnung erhält man die folgenden Relationen :

$$M(\lambda_\kappa, v_\kappa) > \omega_\kappa^2 \iff 2\lambda_\kappa v_\kappa^T v_\kappa + \lambda_\kappa^2 v_\kappa^T A^T A v_\kappa > \lambda_\kappa^2 \omega_\kappa^2 v_\kappa^T v_\kappa \iff$$

$$\lambda_\kappa^2 \left[\omega_\kappa^2 - \frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} \right] < 2\lambda_\kappa$$

Wegen $\lambda_\kappa > 0$ erhält man hieraus die folgende Beziehung :

$$\lambda_\kappa \left[\omega_\kappa^2 - \frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} \right] < 2$$

Daraus ergibt sich unmittelbar die Behauptung des Satzes.

Zusatz :

Für λ_κ lassen sich die beiden folgenden Schranken verwenden :

$$\lambda_\kappa < \frac{2}{\omega_\kappa^2 - \Lambda_1} \quad \text{bzw.} \quad \lambda_\kappa < \frac{2}{\Lambda_n - \Lambda_1}$$

Für die optimale Wahl von λ_κ erhält man :

$$\lambda_\kappa = \frac{\left(\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right) + \sqrt{\left(\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right)^2 + 4 v_\kappa^T v_\kappa}}{2 v_\kappa^T v_\kappa}$$

Die Iteration zur Berechnung von $\max_{\|s\|_3=1} \|As\|_3$ ist nun eindeutig festgelegt. Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.6 :

A sei eine reelle reguläre Matrix. $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ mit $x^{(\mu)T} x^{(\nu)} = \delta_{ij}$ sei ein orthonormiertes System von Eigenvektoren der symmetrischen positiv definiten Matrix $A^T A$ und $0 < \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ seien die zugehörigen Eigenwerte. Der Startvektor s_0 der Iteration besitze die folgende Darstellung :

$$s_0 = \sum_{\nu=1}^{n} c_\nu^{(0)} x^{(\nu)} \quad \text{mit} \quad c_n^{(0)} \neq 0$$

Für den Skalarfaktor λ_κ der Iteration gelte :

$$\lambda_\kappa > 0 \; ; \; 0 < \varepsilon \leq \lambda_\kappa \leq \frac{2}{\left|\frac{v_\kappa^T A^T A v_\kappa}{v_\kappa^T v_\kappa} - \omega_\kappa^2\right|} - \varepsilon \quad \text{mit} \quad \varepsilon \leq \frac{1}{\mu_n}$$

Dann konvergiert die Folge $\omega_\kappa \longrightarrow \sqrt{\mu_n} = \max_{\|s\|_3=1} \|As\|_3$

Beweis :

Die Folge $(\omega_\kappa)_{\kappa=0,1,\ldots}$ ist konvergent, da sie monoton und beschränkt ist. Wegen $0 < \varepsilon \leq \lambda_\kappa$ und nach Konstruktion der Iterationsfolge gilt : $\frac{M'(0, v_\kappa)}{|\lambda_\kappa|} \xrightarrow[\kappa \to \infty]{} 0$. Daraus folgt : 1. $s_\kappa \xrightarrow[\kappa \to \infty]{} \pm x^{(\nu)}$

2. $\omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} \mu_\nu$

Die Behauptung des Satzes lautet : $\mu_\nu = \mu_n$

Wir zeigen dies durch einen indirekten Beweis.

Annahme : $\mu_\nu < \mu_n$

Für die Vektoren s_κ existiert die folgende Darstellung :

$$(*) \qquad s_\kappa = \sum_{\nu=1}^{n} c_\nu^{(\kappa)} x^{(\nu)}$$

Man beachte, daß aus $\mu_\nu < \mu_n$ folgt : $c_n^{(\kappa)} \xrightarrow[\kappa\to\infty]{} 0$

Die Aussage $c_n^{(\kappa)} \xrightarrow[\kappa\to\infty]{} 0$ führen wir zu einem Widerspruch.

Analog wie im Beweis von Satz III.3 erhalten wir die Beziehung :

$$\frac{c_n^{(\kappa+1)}}{c_n^{(\kappa)}} = \frac{1 + \lambda_\kappa(\mu_n - \omega_\kappa^2)}{(1 + \lambda_\kappa^2 v_\kappa^T v_\kappa)^{1/2}}$$

Wegen $\lambda_\kappa > 0$ und $\mu_n - \omega_\kappa^2 > 0$ gilt : $c_n^{(\kappa)} \neq 0$ für $\kappa = 0,1,\ldots$, ausserdem ändert $c_n^{(\kappa)}$ sein Vorzeichen nicht.

Wir zeigen nun $\dfrac{c_n^{(\kappa+1)}}{c_n^{(\kappa)}} > 1$ für $\kappa \geq K$.

Dies ist ein Widerspruch zu $c_n^{(\kappa)} \xrightarrow[\kappa\to\infty]{} 0$

Es gelten - wie unmittelbar ersichtlich - die Relationen:

$$\frac{c_n^{(\kappa+1)}}{c_n^{(\kappa)}} > 1 \iff 1 + \lambda_\kappa^2 v_\kappa^T v_\kappa < (1 + \lambda_\kappa(\mu_n - \omega_\kappa^2))^2 \iff$$

$$\lambda_\kappa^2 (v_\kappa^T v_\kappa - (\mu_n - \omega_\kappa^2)^2) < 2\lambda_\kappa(\mu_n - \omega_\kappa^2)$$

Wegen $\lambda_\kappa > 0$ und $\mu_n - \omega_\kappa^2 > 0$ ist dies äquivalent zu folgender Relation :

$$\lambda_\kappa (v_\kappa^T v_\kappa - (\mu_n - \omega_\kappa^2)^2) < 2(\mu_n - \omega_\kappa^2)$$

Wegen $v_\kappa^T v_\kappa \xrightarrow[\kappa\to\infty]{} 0$ und $\mu_n - \omega_\kappa^2 \geq \mu_n - \mu_\nu$ ist die obige Ungleichung für $\kappa \geq K$ erfüllt, wobei K geeignet zu wählen ist. Dies ist ein Widerspruch zu $c_n^{(\kappa)} \xrightarrow[\kappa\to\infty]{} 0$ und damit zu $\mu_\nu < \mu_n$.

Bemerkung :

Auch hier besteht eine Analogie zu dem bekannten Gradientenverfahren. Der numerische Aufwand bei jedem Iterationsschritt ist genauso groß wie bei der Berechnung von $\min_{\|Sz\|_2=1} \|As\|_2$, allerdings ist die Konvergenzgeschwindigkeit hier im allgemeinen besser.

Auch bei der Berechnung von $\max_{\|s\|_3=1} \|As\|_3$ läßt sich ein weiteres Iterationsverfahren angeben.

Ansatz : $(*) \quad s_{\kappa+1} = \dfrac{s_\kappa + \lambda_\kappa v_\kappa}{\|s_\kappa + \lambda_\kappa v_\kappa\|_3} \quad ; \quad s_\kappa = (s_i^{(\kappa)})$

Wir starten wieder mit einem Einheitsvektor :

$$v_0 := e_{\nu_0} \quad \text{mit} \quad \|Ae_{\nu_0}\|_3 = \max_{\nu=1,\dots,n} \|Ae_\nu\|_3$$

Für den Korrekturvektor v_κ nehmen wir ebenfalls einen Einheitsvektor :

$$v_\kappa := e_{\mu_\kappa}$$

$$\mu_0 = \begin{cases} \nu_0 + 1 & , \text{ falls } \nu_0 < n \\ 1 & , \text{ falls } \nu_0 = n \end{cases} ; \mu_{\kappa+1} = \begin{cases} \mu_\kappa + 1 & , \text{ falls } \mu_\kappa < n \\ 1 & , \text{ falls } \mu_\kappa = n \end{cases}$$

Die Wahl des Skalarfaktors λ_κ wird durch die folgende Forderung festgelegt :

$$M_{\mu_\kappa}(\lambda_\kappa) := \frac{(As_\kappa + \lambda_\kappa a_{\mu_\kappa})^2}{(s_\kappa + \lambda_\kappa e_{\mu_\kappa})^2} \stackrel{!}{=} \max$$

Nach elementarer Rechnung erhält man die folgende Beziehung :

$$M'_{\mu_\kappa}(0) = 2(a_{\mu_\kappa}^T u_\kappa - s_{\mu_\kappa}^{(\kappa)} \omega_\kappa^2)$$

Die Forderung $\omega_{\kappa+1} > \omega_\kappa$ läßt sich genau dann erfüllen, wenn $M'_{\mu_\kappa}(0) \neq 0$.

Durch Auswertung der Gleichung $M'_{\mu_\kappa}(\lambda_\kappa) = 0$ erhält man den optimalen Skalarfaktor λ_κ.

$$\lambda_\kappa = \begin{cases} \frac{1}{2}\left[\left(\frac{\gamma_{\mu_\kappa} - \omega_\kappa}{L_\kappa}\right) + \operatorname{sgn} L_\kappa \sqrt{\left(\frac{\gamma_{\mu_\kappa} - \omega_\kappa}{L_\kappa}\right)^2 + 4\,\frac{S_\kappa}{L_\kappa}}\right] & , \text{ falls } L_\kappa \neq 0 \\ -\dfrac{S_\kappa}{\gamma_{\mu_\kappa} - \omega_\kappa} & , \text{ falls } L_\kappa = 0 \end{cases}$$

Nach diesen Ausführungen ist das Iterationsverfahren eindeutig festgelegt.
Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.7 :

A sei eine reelle reguläre Matrix. $0<\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ seien die Eigenwerte von A^TA . Dann konvergiert die Folge $\omega_\kappa \xrightarrow[\kappa\to\infty]{} \sqrt{\mu_i}$

Beweis :

Der Beweis verläuft analog wie zu Satz III.4 .

Bemerkung :

$(\omega_\kappa)_{\kappa=0,1,\dots}$ konvergiert also nicht notwendig gegen die Wurzel des größten Eigenwertes von A^TA . Experimentelle Untersuchungen haben ergeben, daß die Folge im allgemeinen gegen die Wurzel eines großen Eigenwertes konvergiert. Auch hier liegt die Bedeutung des Verfahrens in einer geeigneten Kombination beider Iterationsverfahren. Über den numerischen Aufwand bei jedem Iterationsschritt gilt das gleiche wie bei der Berechnung von $\min_{\|s\|_3=1}\|As\|_3$. Allerdings ist hier die Konvergenzgeschwindigkeit im allgemeinen grösser als bei der Berechnung von $\min_{\|s\|_3=1}\|As\|_3$.

III. Iterative Berechnung von $\min_{\|s\|_2=1}\|As\|_3$

Wir konstruieren wieder eine geeignete Vektorfolge $(s_\kappa)_{\kappa=0,1,\dots}$

Ansatz : $(*)\quad s_{\kappa+1} = \dfrac{s_\kappa+\lambda_\kappa v_\kappa}{\|s_\kappa+\lambda_\kappa v_\kappa\|_2}\quad ;\quad s_\kappa = (s_i^{(\kappa)})$

Die Iteration ist durch folgende Angaben festgelegt :

1. Angabe eines Startvektors s_0 mit $\|s_0\|_2 = 1$.
2. Angabe eines Korrekturvektors v_κ und eines Skalarfaktors λ_κ bei jedem Iterationsschritt.

Zu 1 : Wir starten mit einem Einheitsvektor :

$$s_0 := e_{\nu_0} \quad \text{mit} \quad \|Ae_{\nu_0}\|_3 = \min_{\nu=1,\dots,n} \|Ae_\nu\|_3$$

Zu 2 : Für den Korrekturvektor v_κ wählen wir ebenfalls einen Einheitsvektor :

$$v_\kappa := e_{\mu_\kappa}$$

$$\mu_0 := \begin{cases} \nu_0 + 1 & , \text{ falls } \nu_0 < n \\ 1 & , \text{ falls } \nu_0 = n \end{cases} \quad ; \mu_{\kappa+1} = \begin{cases} \mu_\kappa + 1 & , \text{ falls } \mu_\kappa < n \\ 1 & , \text{ falls } \mu_\kappa = n \end{cases}$$

Wir ändern also bei jedem Iterationsschritt genau eine Komponente des Vektors s_κ und führen anschließend die Normierung durch.

Bei der Wahl des Skalarfaktors λ_κ beachten wir die folgende Forderung :

$$\|As_{\kappa+1}\|_3 = \frac{\|As_\kappa + \lambda_\kappa a_{\mu_\kappa}\|_3}{\|s_\kappa + \lambda_\kappa e_{\mu_\kappa}\|_2} < \|As_\kappa\|_3$$

Wir führen die folgenden Bezeichnungen ein :

$$u_\kappa := As_\kappa \quad ; \quad \omega_\kappa := \|u_\kappa\|_3$$

Für die weiteren Betrachtungen ist es günstig, von der folgenden äquivalenten Forderung auszugehen :

$$M_{\mu_\kappa}(\lambda_\kappa) := \frac{(u_\kappa + \lambda_\kappa a_{\mu_\kappa})^2}{(\|s + \lambda_\kappa e_{\mu_\kappa}\|_2)^2} = \omega_{\kappa+1}^2 < (As_\kappa)^2 = \omega_\kappa^2$$

Man beachte : $\|s_\kappa + \lambda_\kappa e_{\mu_\kappa}\|_2 = \sum_{\nu \neq \mu_\kappa} |s_\nu^{(\kappa)}| + |s_{\mu_\kappa}^{(\kappa)} + \lambda_\kappa|$

Hieraus ergeben sich unmittelbar die folgenden Relationen :

I. $s_{\mu_\kappa}^{(\kappa)} > 0$ Daraus folgt : $\|s_\kappa + \lambda_\kappa e_{\mu_\kappa}\|_2 = 1 + \lambda_\kappa$ für $\lambda_\kappa \geq -s_{\mu_\kappa}^{(\kappa)}$

II. $s_{\mu_\kappa}^{(\kappa)} < 0$ Daraus folgt : $\|s_\kappa + \lambda_\kappa e_{\mu_\kappa}\|_2 = 1 - \lambda_\kappa$ für $\lambda_\kappa \leq -s_{\mu_\kappa}^{(\kappa)}$

III. $s_{\mu_\kappa}^{(\kappa)} = 0$ Daraus folgt : $\|s_\kappa + \lambda_\kappa e_{\mu_\kappa}\|_2 = \begin{cases} 1 + \lambda_\kappa \text{ für } \lambda_\kappa \geq 0 \\ 1 - \lambda_\kappa \text{ für } \lambda_\kappa \leq 0 \end{cases}$

Für die Diskussion der Funktion $M_{\mu_\kappa}(\lambda_\kappa)$ empfiehlt es sich, die beiden folgenden Hilfsfunktionen einzuführen :

$$M_{1,\mu_\kappa}(\lambda_\kappa) := \frac{(As_\kappa+\lambda_\kappa a_{\mu_\kappa})^2}{(1+\lambda_\kappa)^2} \quad ; \quad M_{2,\mu_\kappa}(\lambda_\kappa) := \frac{(As_\kappa+\lambda_\kappa a_{\mu_\kappa})^2}{(1-\lambda_\kappa)^2}$$

Nach obigen Ausführungen gilt die folgende Beziehung :

$$M_{\mu_\kappa}(\lambda_\kappa) = \begin{cases} M_{1,\mu_\kappa}(\lambda_\kappa) & \text{für } s_{\mu_\kappa}^{(\kappa)} \geq 0 \text{ und } \lambda_\kappa \geq -s_{\mu_\kappa}^{(\kappa)} \\ M_{2,\mu_\kappa}(\lambda_\kappa) & \text{für } s_{\mu_\kappa}^{(\kappa)} \leq 0 \text{ und } \lambda_\kappa \leq -s_{\mu_\kappa}^{(\kappa)} \end{cases}$$

Nach elementarer Rechnung erhält man die folgenden Beziehungen :

$$M_{1,\mu_\kappa}'(0) = 2(a_{\mu_\kappa}^T u_\kappa - \omega_\kappa^2) \quad ; \quad M_{2,\mu_\kappa}'(0) = 2(a_{\mu_\kappa}^T u_\kappa + \omega_\kappa^2)$$

Folgerung :

I. Falls $s_{\mu_\kappa}^{(\kappa)} > 0$ und $M_{1,\mu_\kappa}'(0) \neq 0$, so existiert stets ein Faktor λ_κ , so daß gilt : $\omega_{\kappa+1} < \omega_\kappa$

II. Falls $s_{\mu_\kappa}^{(\kappa)} < 0$ und $M_{2,\mu_\kappa}'(0) \neq 0$, so existiert stets ein Faktor λ_κ , so daß gilt : $\omega_{\kappa+1} < \omega_\kappa$

III. Falls $s_{\mu_\kappa}^{(\kappa)} = 0$, so existiert stets ein Faktor λ_κ , so daß gilt : $\omega_{\kappa+1} < \omega_\kappa$. Man beachte dabei, daß stets mindestens eine der beiden folgenden Relationen gilt : $M_{1,\mu_\kappa}'(0) < 0$; $M_{2,\mu_\kappa}'(0) > 0$.

Durch die Forderung $M_{1,\mu_\kappa}'(\lambda_\kappa) \stackrel{!}{=} \min$ bzw. $M_{2,\mu_\kappa}(\lambda_\kappa) \stackrel{!}{=} \min$ erhält man die folgenden Werte :

$$\lambda_\kappa^{(1)} = \frac{1}{2}\left[\left(\frac{a_{\mu_\kappa}^T a_{\mu_\kappa} - \omega_\kappa^2}{a_{\mu_\kappa}^T u_\kappa - a_{\mu_\kappa}^T a_{\mu_\kappa}}\right) + \sqrt{\left(\frac{a_{\mu_\kappa}^T a_{\mu_\kappa} - \omega_\kappa^2}{a_{\mu_\kappa}^T u_\kappa - a_{\mu_\kappa}^T a_{\mu_\kappa}}\right)^2 - 4\left(\frac{a_{\mu_\kappa}^T u_\kappa - \omega_\kappa^2}{a_{\mu_\kappa}^T a_{\mu_\kappa} - a_{\mu_\kappa}^T u_\kappa}\right)}\right]$$

$$\lambda_\kappa^{(2)} = \frac{1}{2}\left[\left(\frac{a_{\mu_\kappa}^T a_{\mu_\kappa} - \omega_\kappa^2}{a_{\mu_\kappa}^T u_\kappa + a_{\mu_\kappa}^T a_{\mu_\kappa}}\right) - \sqrt{\left(\frac{a_{\mu_\kappa}^T a_{\mu_\kappa} - \omega_\kappa^2}{a_{\mu_\kappa}^T u_\kappa + a_{\mu_\kappa}^T a_{\mu_\kappa}}\right)^2 + 4\left(\frac{a_{\mu_\kappa}^T u_\kappa + \omega_\kappa^2}{a_{\mu_\kappa}^T a_{\mu_\kappa} + a_{\mu_\kappa}^T u_\kappa}\right)}\right]$$

Bei der Wahl des Faktors λ_κ ist für $s_{\mu_\kappa}^{(\kappa)} \neq 0$ darauf zu achten, daß $s_{\mu_\kappa}^{(\kappa+1)}$ nicht sein Vorzeichen ändert. Die folgende Vorschrift bestimmt eindeutig den Faktor λ_κ.

I. $s_{\mu_\kappa}^{(\kappa)} > 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(1)}, & \text{falls } \lambda_\kappa^{(1)} \geq -s_{\mu_\kappa}^{(\kappa)} \\ -s_{\mu_\kappa}^{(\kappa)} & \text{sonst} \end{cases}$

II. $s_{\mu_\kappa}^{(\kappa)} < 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(2)}, & \text{falls } \lambda_\kappa \leq -s_{\mu_\kappa}^{(\kappa)} \\ -s_{\mu_\kappa}^{(\kappa)} & \text{sonst} \end{cases}$

III. $s_{\mu_\kappa}^{(\kappa)} = 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(1)}, & \text{falls } a_{\mu_\kappa}^T u_\kappa \leq 0 \\ \lambda_\kappa^{(2)}, & \text{falls } a_{\mu_\kappa}^T u_\kappa > 0 \end{cases}$

Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.8 :

A sei eine reelle reguläre Matrix. Dann existiert ein Vektor $e \in \mathcal{E}$, so daß gilt : $\omega_\kappa \xrightarrow[\kappa \to \infty]{} \frac{1}{\|A^{-1T} e\|_3}$

Beweis :

I. Die Folge $(\omega_\kappa)_{\kappa = 0,1,\ldots}$ ist konvergent, da sie monoton und beschränkt ist.

II. Nach Konstruktion der Iteration gilt für $\nu = 1,2,\ldots,n$ entweder $a_\nu^T u_\kappa - \omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} 0$ oder $a_\nu^T u_\kappa + \omega_\kappa^2 \xrightarrow[\kappa \to \infty]{} 0$.

Es existiert also ein Vektor $e \in \mathcal{E}$ mit $A^T u_\kappa - \omega_\kappa^2 e \xrightarrow[\kappa \to \infty]{} 0$.

Daraus folgt : $u_\kappa \xrightarrow[\kappa \to \infty]{} \omega_\kappa^2 A^{-1T} e$. Daraus folgt :

$\omega_\kappa = \|u_\kappa\|_3 \xrightarrow[\kappa \to \infty]{} \omega_\kappa^2 \|A^{-1T} e\|_3$. Daraus folgt :

$\omega_\kappa \xrightarrow[\kappa \to \infty]{} \frac{1}{\|A^{-1T} e\|_3}$

Zusatz :

Wendet man die angegebene Iteration auf A^T an, so folgt unmittelbar $\omega_\kappa \xrightarrow[\kappa \to \infty]{} \frac{1}{\|A^{-1} e\|_3}$

Bemerkung :

Der im Grenzwert auftretende Vektor e wird im allgemeinen nicht der optimale Vektor e_0 mit $\|A^{-1}e_0\|_2 = \max_{e \in \mathcal{E}} \|A^{-1}e\|_2$ sein. Experimentelle Untersuchungen haben ergeben, daß sich der Grenzwert der Folge im allgemeinen als guter Näherungswert für den optimalen Wert eignet.

Numerischer Aufwand der Iteration :

Nach Berechnung der Werte $a_\nu^T a_\nu$, $\nu = 1, 2, \ldots, n$ sind bei jedem Iterationsschritt im wesentlichen die beiden Skalarprodukte $a_\kappa^T u_\kappa$ und $u_\kappa^T u_\kappa$ zu berechnen. Die Anzahl der Iterationen hängt wesentlich ab von dem Startvektor und der Struktur der Matrix A. Das Verfahren hat den Vorteil eines geringen numerischen Aufwandes.

IV. Iterative Berechnung von $\|A^{-1}\|_1$ und $\|A^{-1}\|_2$.

Bei der Berechnung von $\|A^{-1}\|_1$ bzw. $\|A^{-1}\|_2$ benötigt man nicht die gesamte inverse Matrix, sondern es genügt die Kenntnis eines geeigneten Zeilen- bzw. Spaltenvektors.

Bei der Invertierung sehr großer Matrizen berechnet man häufig zunächst eine Näherungsmatrix $B \approx A^{-1}$, die man dann iterativ verbessert. Diese iterative Verbesserung ist mit sehr großem numerischen Aufwand verbunden. Die Idee des folgenden Iterationsverfahrens besteht nun darin, daß man nicht die gesamte Matrix B iterativ verbessert, sondern nur den Zeilen- bzw. Spaltenvektor von B mit maximaler Betragssummennorm. Das Iterationsverfahren konvergiert gegen den entsprechenden Zeilen- bzw. Spaltenvektor von A^{-1}, der dann zumindest einen Näherungswert für $\|A^{-1}\|_1$ bzw. $\|A^{-1}\|_2$ liefert.

Den weiteren Ausführungen legen wir folgende Bezeichnungen zugrunde : $A = (a_1\ a_2\ \ldots\ a_n)$; $A^{-1} = (s_1\ s_2\ \ldots\ s_n) = (z_\nu^T)$.

Es gilt dann : $\|A^{-1}\|_1 = \max_{\nu=1,\ldots,n} \|z_\nu\|_2 = \|z_{\nu_0}\|_2$; $\|A^{-1}\|_2 = \max_{\mu=1,\ldots,n} \|s_\mu\|_2 = \|s_{\mu_0}\|_2$

Bei der Wahl des Faktors λ_κ ist für $s_{\mu_\kappa}^{(\kappa)} \neq 0$ darauf zu achten, daß $s_{\mu_\kappa}^{(\kappa+\nu)}$ nicht sein Vorzeichen ändert. Die folgende Vorschrift bestimmt eindeutig den Faktor λ_κ.

I. $s_{\mu_\kappa}^{(\kappa)} > 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(1)}, & \text{falls } \lambda_\kappa^{(1)} \geq -s_{\mu_\kappa}^{(\kappa)} \\ -s_{\mu_\kappa}^{(\kappa)} & \text{sonst} \end{cases}$

II. $s_{\mu_\kappa}^{(\kappa)} < 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(2)}, & \text{falls } \lambda_\kappa \leq -s_{\mu_\kappa}^{(\kappa)} \\ -s_{\mu_\kappa}^{(\kappa)} & \text{sonst} \end{cases}$

III. $s_{\mu_\kappa}^{(\kappa)} = 0$

Man wähle $\lambda_\kappa = \begin{cases} \lambda_\kappa^{(1)}, & \text{falls } a_{\mu_\kappa}^T u_\kappa \leq 0 \\ \lambda_\kappa^{(2)}, & \text{falls } a_{\mu_\kappa}^T u_\kappa > 0 \end{cases}$

Der folgende Satz gibt Auskunft über das Konvergenzverhalten der Iteration.

Satz III.8 :

A sei eine reelle reguläre Matrix. Dann existiert ein Vektor $e \in \mathcal{E}$, so daß gilt : $\omega_\kappa \xrightarrow[\kappa\to\infty]{} \frac{1}{\|A^{-1T}e\|_3}$

Beweis :

I. Die Folge $(\omega_\kappa)_{\kappa=0,1,\ldots}$ ist konvergent, da sie monoton und beschränkt ist.

II. Nach Konstruktion der Iteration gilt für $\nu = 1,2,\ldots,n$ entweder $a_\nu^T u_\kappa - \omega_\kappa^2 \xrightarrow[\kappa\to\infty]{} 0$ oder $a_\nu^T u_\kappa + \omega_\kappa^2 \xrightarrow[\kappa\to\infty]{} 0$.

Es existiert also ein Vektor $e \in \mathcal{E}$ mit $A^T u_\kappa - \omega_\kappa^2 e \xrightarrow[\kappa\to\infty]{} 0$.

Daraus folgt : $u_\kappa \xrightarrow[\kappa\to\infty]{} \omega_\kappa^2 A^{-1T} e$. Daraus folgt :

$\omega_\kappa = \|u_\kappa\|_3 \xrightarrow[\kappa\to\infty]{} \omega_\kappa^2 \|A^{-1T} e\|_3$. Daraus folgt :

$\omega_\kappa \xrightarrow[\kappa\to\infty]{} \frac{1}{\|A^{-1T} e\|_3}$

Zusatz :

Wendet man die angegebene Iteration auf A^T an, so folgt unmittelbar $\omega_\kappa \xrightarrow[\kappa\to\infty]{} \frac{1}{\|A^{-1} e\|_3}$

Bemerkung :

Der im Grenzwert auftretende Vektor e wird im allgemeinen nicht der optimale Vektor e_o mit $\| A^{-1} e_o \|_3 = \max_{e \in \mathcal{E}} \| A^{-1} e \|_3$ sein. Experimentelle Untersuchungen haben ergeben, daß sich der Grenzwert der Folge im allgemeinen als guter Näherungswert für den optimalen Wert eignet.

Numerischer Aufwand der Iteration :

Nach Berechnung der Werte $a_\nu^T a_\nu$ $\nu = 1,2,\ldots,n$ sind bei jedem Iterationsschritt im wesentlichen die beiden Skalarprodukte $a_\kappa^T u_\kappa$ und $u_\mu^T u_\kappa$ zu berechnen. Die Anzahl der Iterationen hängt wesentlich ab von dem Startvektor und der Struktur der Matrix A. Das Verfahren hat den Vorteil eines geringen numerischen Aufwandes.

IV. Iterative Berechnung von $\| A^{-1} \|_1$ und $\| A^{-1} \|_2$.

Bei der Berechnung von $\| A^{-1} \|_1$ bzw. $\| A^{-1} \|_2$ benötigt man nicht die gesamte inverse Matrix, sondern es genügt die Kenntnis eines geeigneten Zeilen- bzw. Spaltenvektors.

Bei der Invertierung sehr großer Matrizen berechnet man häufig zunächst eine Näherungsmatrix $B \approx A^{-1}$, die man dann iterativ verbessert. Diese iterative Verbesserung ist mit sehr großem numerischen Aufwand verbunden. Die Idee des folgenden Iterationsverfahrens besteht nun darin, daß man nicht die gesamte Matrix B iterativ verbessert, sondern nur den Zeilen- bzw. Spaltenvektor von B mit maximaler Betragssummennorm. Das Iterationsverfahren konvergiert gegen den entsprechenden Zeilen- bzw. Spaltenvektor von A^{-1}, der dann zumindest einen Näherungswert für $\| A^{-1} \|_1$ bzw. $\| A^{-1} \|_2$ liefert.

Den weiteren Ausführungen legen wir folgende Bezeichnungen zugrunde : $A = (a_1\ a_2\ \ldots\ a_n)$; $A^{-1} = (s_1\ s_2\ \ldots\ s_n) = (z_\nu^T)$.

Es gilt dann : $\| A^{-1} \|_1 = \max_{\nu=1,\ldots,n} \| z_\nu \|_2 = \| z_{\nu_o} \|_2$; $\| A^{-1} \|_2 = \max_{\mu=1,\ldots,n} \| s_\mu \|_2 = \| s_{\mu_o} \|_2$

Wir werden uns auf die Berechnung von $\| A^{-1} \|_1$ konzentrieren. Die Berechnung von $\| A^{-1} \|_2$ erfolgt dann ganz analog. Zunächst sind einige vorbereitende Betrachtungen notwendig.

Hilfssatz III.2 :

Gegeben seien die reelle reguläre Matrix A, der Index $1 \leq \nu_0 \leq n$ und die reelle Zahl $\alpha \neq 0$.

$$\mathcal{T} := \left\{ t = (t_\nu) \mid t_{\nu_0} = \alpha \right\}$$

Dann existiert genau ein Vektor $t^* \in \mathcal{T}$ mit der Eigenschaft :

$$\| At^* \|_3 = \min_{t \in \mathcal{T}} \| At \|_3$$

Für diesen Vektor t^* gilt die folgende Beziehung :

$$At^* = \mu z_{\nu_0} \quad \text{mit} \quad \mu = a_{\nu_0}^T At^* \neq 0$$

Beweis :

Für den Vektor z_{ν_0} existiert die folgende Darstellung :

$$z_{\nu_0} = \sum_{\nu=1}^{n} \sigma_\nu a_\nu$$

Es gilt : $z_{\nu_0}^T z_{\nu_0} = \sum_{\nu=1}^{n} \sigma_\nu z_{\nu_0}^T a_\nu = \sigma_{\nu_0} \neq 0$

Wir setzen $t_0 = (t_\nu^{(0)})$ mit $t_\nu^{(0)} = \dfrac{\alpha}{\sigma_{\nu_0}} \sigma_\nu$

Dann gilt : $t_0 \in \mathcal{T}$; $At_0 = \sum_{\nu=1}^{n} t_\nu^{(0)} a_\nu = \dfrac{\alpha}{\sigma_{\nu_0}} z_{\nu_0} = \mu z_{\nu_0}$

Für $t_0 \neq t = (t_\nu) \in \mathcal{T}$ gilt die folgende Darstellung :

$$t = t_0 + v \quad ; \quad v = (v_\nu) \quad \text{mit} \quad v_{\nu_0} = 0$$

Daraus folgt : $z_{\nu_0}^T Av = z_{\nu_0}^T \sum_{\nu \neq \nu_0} v_\nu a_\nu = \sum_{\nu \neq \nu_0} v_\nu z_{\nu_0}^T a_\nu = 0$

Wir setzen $u := At$; $u_0 := At_0$; $z := Av$

Daraus folgt : $u = u_0 + z$; $z^T u_0 = 0$

Daraus folgt : $u^T u = u_0^T u_0 + z^T z > u_0^T u_0$

Daraus folgt : $\| At \|_3 > \| At_0 \|_3 \qquad t^* := t_0$

Man beachte $a_{\nu_0}^T At^* = \mu a_{\nu_0}^T z_{\nu_0} = \mu$ q.e.d.

Wir konstruieren nun eine Vektorfolge $(t_\kappa)_{\kappa = 0,1,\ldots}$ $t_\kappa \in \mathcal{T}$, welche gegen t^* konvergiert.

Ansatz : $t_{\kappa+1} = t_\kappa + \lambda_\kappa e_{\mu_\kappa}$

Wir starten mit einem beliebigen Vektor $t_0 \in \mathcal{T}$, z.B. $t_0 = (t_i^{(0)})$,

wobei gilt : $t_\nu^{(0)} = \begin{cases} 0 & \text{für } \nu \neq \nu_0 \\ \alpha & \text{für } \nu = \nu_0 \end{cases}$

μ_κ durchlaufe zyklisch die Indexmenge $\{1,2,\ldots,n\} - \{\nu_0\}$

Für den Skalarfaktor λ_κ wähle man $\lambda_\kappa = -\dfrac{a_{\mu_\kappa} A t_\kappa}{a_{\mu_\kappa}^T a_{\mu_\kappa}}$

Dadurch ist die Iteration eindeutig festgelegt.

Hilfssatz III.3 :

Die angegebene Iteration $t_{\kappa+1} = t_\kappa + \lambda_\kappa e_{\mu_\kappa}$ konvergiert gegen den Vektor t^* .

Beweis :

Wir setzen $M_{\mu_\kappa}(\lambda_\kappa) := t_{\kappa+1}^T A^T A t_{\kappa+1} = t_\kappa^T A^T A t_\kappa + 2\lambda_\kappa a_{\mu_\kappa}^T A t_\kappa + \lambda_\kappa^2 a_{\mu_\kappa}^T a_{\mu_\kappa}$

Forderung : $M_{\mu_\kappa}(\lambda_\kappa) \stackrel{!}{=} \min$

Durch Auswertung der Gleichung $M'_{\mu_\kappa}(\lambda_\kappa) = 0$ erhält man :

$$\lambda_\kappa = -\frac{a_{\mu_\kappa}^T A t_\kappa}{a_{\mu_\kappa}^T a_{\mu_\kappa}}$$

Die Wahl des Faktors λ_κ ist also hinsichtlich der Minimierung von $\|At\|_3$ optimal.

Nach Konstruktion der Iteration gilt die folgende Beziehung :

$$M'_{\mu_\kappa}(0) = 2a_{\mu_\kappa}^T A t_\kappa \xrightarrow[\kappa\to\infty]{} 0$$

Daraus folgt : $a_\nu^T A t_\kappa \xrightarrow[\kappa\to\infty]{} 0$ für $\nu \neq \nu_0$

Daraus folgt : $A t_\kappa \xrightarrow[\kappa\to\infty]{} \mu z_{\nu_0}$ Daraus folgt : $t_\kappa \xrightarrow[\kappa\to\infty]{} t^*$

Zusatz :

Setzt man $u_\kappa := A t_\kappa$, so erhält man die konvergente Folge $(u_\kappa)_{\kappa=0,1,\ldots}$

Es gilt die Iteration : $u_{\kappa+1} = u_\kappa + \lambda_\kappa a_{\mu_\kappa}$

Satz III.9 :

Gegeben seien die reelle reguläre Matrix A und der Index $1 \le \nu_0 \le n$

$$u_0 = \sum_{\nu=1}^{n} \tau_\nu a_\nu \quad \text{mit} \quad \tau_{\nu_0} \neq 0$$

sei ein beliebiger Startvektor für die folgende Iteration :

$$u_{\kappa+1} = u_\kappa + \lambda_\kappa a_{\mu_\kappa}$$

Dabei durchlaufe μ_κ zyklisch die Indexmenge $\{1,2,\ldots,n\} - \{\nu_0\}$.

$$\lambda_\kappa = -\frac{a_{\mu_\kappa}^T u_\kappa}{a_{\mu_\kappa}^T a_{\mu_\kappa}}$$

Dann gilt: $u_\kappa \xrightarrow[\kappa\to\infty]{} u^* = \mu z_{\nu_0}$; $\mu = a_{\nu_0}^T u^*$

Beweis :

$$\gamma := \left\{ t = (t_\nu) \mid t_{\nu_0} = \tau_{\nu_0} \right\}$$

Wir betrachten die folgende Iteration :

$$t_{\kappa+1} = t_\kappa + \lambda_\kappa e_{\mu_\kappa}$$

Ein Vektor $t_0 = (t_\nu^{(0)})$ mit $t_\nu^{(0)} = \tau_\nu$ sei der Startvektor der Iteration.

Nach Hilfssatz III.3 und dem Zusatz folgt unmittelbar die Behauptung des Satzes.

Numerischer Aufwand der Iteration :

Nach Berechnung der Größen $a_\nu^T a_\nu$ $\nu = 1,2,\ldots,n$ ist bei jedem Iterationsschritt im wesentlichen das Skalarprodukt $a_{\mu_\kappa}^T u_\kappa$ zu berechnen. Die Anzahl der Iterationen hängt wesentlich von dem Startvektor ab. Das Verfahren hat den Vorteil eines geringen numerischen Aufwandes.

Praktische Anwendung :

Man benötigt zunächst einen Index ν_0, dessen zugehöriger Zeilen- bzw. Spaltenvektor von A^{-1} die Norm $\|A^{-1}\|_1$ bzw. $\|A^{-1}\|_2$ exakt oder zumindest näherungsweise liefert. Diesen Index ν_0 kann man etwa mit Hilfe einer groben Näherungsmatrix $B \approx A^{-1}$ gewinnen.

Berechnungsschema :

1. Berechnung einer Näherungsmatrix $B \approx A^{-1}$.
2. Man suche den Zeilenvektor b_{ν_0} von B mit maximaler Betragssummennorm.
3. Man iteriere $u_{\kappa+1} = u_\kappa + \lambda_\kappa a_{\mu_\kappa}$ mit $u_0 = b_{\nu_0}$.
4. Beim Index $\kappa = K_0$ breche man die Iteration ab.

Dann gilt: $z_{\nu_0} \approx (a_{\nu_0}^T u_{\kappa_0}) u_{\kappa_0}$; $\|A^{-1}\|_1 \approx |a_{\nu_0} u_{\kappa_0}| \|u_{\kappa_0}\|_2$

Bemerkung :

1. Wendet man das Iterationsverfahren auf A^T statt A an, so erhält man eine Näherung für $\|A^{-1}\|_2$.
2. Die Lösung des Gleichungssystems $A^T x = e_\nu$ mit $e_\nu^T = (0,..,\overset{\nu}{1},..,0)$ ist der ν-te Zeilenvektor von A. Das angegebene Iterationsverfahren kann man daher auch auffassen als Lösungsverfahren für das obige spezielle Gleichungssystem.

Für alle Iterationsverfahren, die in Kapitel III § 2 angegeben wurden, sind im Kapitel IV Algolprogramme angefertigt worden. Mit Hilfe dieser Algolprogramme lassen sich die erzielten Ergebnisse testen.

IV. KAPITEL.

Algolprogramme

§ 1. Allgemeine Bemerkungen

I. Programmsprache

Die vorliegenden Programme sind in Algol 60 geschrieben. Da dort noch keine Ein- und Ausgabeprozeduren festgelegt sind, sind diese im allgemeinen von Rechenanlage zu Rechenanlage verschieden. Die vorliegende Fassung bezieht sich auf die Rechenanlage EL-X8 im Rechenzentrum der Universität Würzburg. Eingabegerät für Programm und Daten ist der Lochstreifenleser EL-1000; die Ausgabe erfolgt auf dem Schnelldrucker; vom Programm benutzt wird der ganze 32K-Kernspeicher der Anlage. Das Programm ist im 5-Kanal-Telecode abgelocht. Die Daten sind für die vorliegenden Programme auf Lochstreifen nach der READ-Konvention der ALCOR-Gruppe abzulochen.

II. Bibliotheksprogramme

Aus der Würzburger Programmbibliothek wurden die folgenden Programme verwendet:

1. Die Prozedur INVERS zur Invertierung einer Matrix.
2. Die Prozedur MJACOBI zur Berechnung der Eigenwerte einer reellen symmetrischen Matrix.
3. Die Prozedur COPIV zur Lösung eines linearen Gleichungssystems.
4. Die Prozedur RANDOM zur Erzeugung von gleichverteilten Zufallszahlen.

Eine genaue Beschreibung dieser Prozeduren befindet sich in der Würzburger Programmbibliothek.

§ 2. Algolprogramme mit Diskussion

I. Algolprogramm zur Untersuchung zufälliger Fehler in den Ausgangsdaten.

1. Bemerkungen zum Programm

Es wurden die Prozeduren RANDOM und COPIV verwendet. Die Ein- und Ausgabedaten sind unmittelbar dem Programm zu entnehmen.

2. Programm

```
'begin'

'procedure'copiv(a,b,n,m,eps,det,sing);
'value'n,m,eps;
'real'eps,det;
'integer'n,m,sing;
'array'a,b;
'begin'
'real'c,sum;
'integer'i,j,k,ri,ci,con;
'integer''array's[1:n];
'for'i:=1'step'1'until'n'do's[i]:=i;
'for'i:=1'step'1'until'n'do'
'begin' ri:=ci:=i;
  'for'j:=i'step'1'until'n'do'
  'begin''for'k:=i'step'1'until'n'do'
    'begin''if'abs(a[j,k])'greater'abs(a[ri,ci])'then'
      'begin'ri:=j;
        ci:=k;
      'end';
    'end';
  'end';
  'if'ri'notequal'i'then'
  'begin''for'j:=1'step'1'until'n'do'
    'begin'c:=a[i,j];
      a[i,j]:=a[ri,j];
      a[ri,j]:=c;
    'end';
    'for'j:=1'step'1'until'm'do'
    'begin'c:=b[i,j];
      b[i,j]:=b[ri,j];
      b[ri,j]:=c;
    'end';
  'end';
  'if'ci'notequal'i'then'
  'begin''for'j:=1'step'1'until'n'do'
    'begin'c:=a[j,i];
      a[j,i]:=a[j,ci];
      a[j,ci]:=c;
    'end';
  'end';
  con:=s[i];
  s[i]:=s[ci];
  s[ci]:=con;
  'if'abs(a[i,i])'less'eps'then'
  'begin'
  sing:=1;
  c:=0;
  'go to' se;
  'end';
  'if'i'equal'n'then''goto'sol;
  'for'j:=i+1'step'1'until'n'do'
  'begin'c:=a[j,i]/a[i,i];
    'for'k:=i+1'step'1'until'n'do'a[j,k]:=a[j,k]-c*a[i,k];
    'for'k:=1'step'1'until'm'do'b[j,k]:=b[j,k]-c*b[i,k];
    a[j,i]:=c;
  'end';
```

```
'end';

sol: 'for'k:=1'step'1'until'm'do'b[n,k]:=b[n,k]/a[n,n];
'for'i:=n-1'step'-1'until'1'do'
'begin''for'j:=1'step'1'until'm'do'
   'begin'sum:=0;
      'for'k:=i+1'step'1'until'n'do'sum:=sum+a[i,k]*b[k,j];
      b[i,j]:=(b[i,j]-sum)/a[i,i];
   'end';
'end';
c:=1;
'for'i:=1'step'1'until'm'do'
'begin'c:=c*a[i,i];
   'for'k:=1'step'1'until'n'do'
   'if'a[k]'equal'i'then''goto'e;
e: 'for'j:=1'step'1'until'm'do'
   'begin'sum:=b[k,j];
      b[k,j]:=b[i,j];
      b[i,j]:=sum;
   'end';
      con:=s[k];
      s[k]:=s[i];
      s[i]:=con;
'end';
sing:=0;
se: det:=c;
'end';

'real' 'procedure' random(first);
       'boolean' first;
       'begin'
'integer' t, tdi;
'own' 'integer' i, p;
'real' rd;
'if' first 'then' 'begin' i := 2; p := -1; 'end'
'else' i := i + 1;
l:
t := abs(time);
'if' i 'notless' t/6000  'then' i := 2;
tdi := t'div'i;
'if' tdi 'equal' p 'then' 'begin' i := 2*i; 'go to' l; 'end';
random := t/i - tdi;
p := tdi;
          'end' random;
'begin'
'integer' n;
n:=10;
mk:   'begin'
'integer'i,j,sing;
'real' bb,eps,det,c1,c2,c3,k,cc1,cc2,cc3;
'array' a[1:n,1:n],aa[1:n,1:n],b[1:n],ab[1:n,1:1],x[1:n],y[1:n];
bb:=random('true');
'for'i:=1 'step' 1 'until' 50 'do' bb:=random('false');
```

```
'for' i:=1 'step' 1 'until' n 'do' 'begin'
b[i]:=(random('false')-0.5)×20;
ab[i,1]:=b[i];
'for' j:=1 'step' 1 'until' n 'do' 'begin'
a[i,j]:=(random('false')-0.5)×20;
aa[i,j]:=a[i,j]; 'end'; 'end';
eps:=1₁₀-11; det:=0; sing:=0;
copiv(aa,ab,n,1,eps,det,sing);
'for' i:=1 'step' 1 'until' n 'do' x[i]:=ab[i,1];
k:=1/100;

mkk:   'for' i:=1 'step' 1 'until' n 'do' 'begin'
ab[i,1]:=b[i];
'for'j:=1 'step' 1 'until' n 'do' 'begin'
bb:=(random('false')-0.5)×20×k;
aa[i,j]:=a[i,j]+bb; 'end'; 'end';
copiv(aa,ab,n,1,eps,det,sing);
'for' i:=1 'step' 1 'until' n 'do' y[i]:=ab[i,1];
nlcr; write(''eps/m=''); fixt(1,3,k); nlcr;
c1:=0; c2:=0; c3:=0; cc1:=0; cc2:=0; cc3:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
bb:=abs(y[i]-x[i]); det:=abs(x[i]);
'if' bb 'greater' c1 'then' c1:=bb;
'if' det 'greater' cc1 'then' cc1:=det;
c2:=c2+bb; cc2:=cc2+det;
c3:=c3+bb×bb; cc3:=cc3+det×det; 'end';
'if' cc1 'less' 1₁₀-600 'then' 'begin'
write(''normenim meanner sind 0''); 'goto' aa; 'end';
c1:=c1/cc1; c2:=c2/cc2; c3:=sqrt(c3/cc3);

nlcr; write(''maximumsnorm von y-x/x'');
flot(13,3,c1); nlcr;
write(''betragsnorm von y-x/x'');
flot(13,3,c2); nlcr;
write(''euklidische norm von y-x/x'');
flot(13,3,c3); nlcr; nlcr;
aa:   'if' k 'less' 1/20 'then' 'begin' k:=1/10; 'goto' mkk; 'end';
'if' k 'less' 1/5 'then' 'begin' k:=1/2; 'goto' mkk; 'end';
'if' n 'equal' 10 'then' 'begin' n:=20; 'goto' ak; 'end';
'if' n 'equal' 20 'then' 'begin' n:=50; 'goto' ak; 'end';
'end'; 'end'; 'end';⊙
```

II. Algolprogramm zur exakten und näherungsweisen Berechnung der Größen α_i.

1. Bemerkungen zu den Prozeduren im Programm.

a) Die Prozedur LINES(P) bewirkt P Zeilenverschube auf dem Schnelldrucker.

b) Die Prozedur OUTPUT(MAT,N) bewirkt das Ausdrucken von quadratischen Matrizen, die im 'ARRAY' MAT [1:N,1:N] stehen, sodaß jede Matrizenzeile als Block von Zeilen aus je 10 Matrixelementen ausgegeben wird.

c) Die Prozedur ATA(MA,MB,N) liefert das Produkt der zur Matrix MA transponierten Matrix MA^T mit MA in MB ab.

d) Die Prozedur DAD(BI,N) wirkt auf einen Vektor BI [1:N] , dessen Elemente nur -1 oder +1 sind. Sie erzeugt, indem -1 als 0 und +1 als L aufgefaßt werden, durch Dualzahladdition eine neue Kombination von +1 und -1 in BI.

e) Die Prozedur MAXNO(A,N,M,P,V,NORM,F) liefert eine maximale Vektornorm in NORM und den dazugehörigen maximalen Vektor in V [1:N] ab. Dabei kann eine beliebige Normfunktion F verwendet werden. Die Berechnung des Vektors V erfolgt aus der Matrix A [1:N,1:M] durch Multiplikation mit einem Vorzeichenvektor der Dimension M. P ist die Nummer der zugrundegelegten Norm und tritt als Parameter von F auf. Der Vorzeichenvektor wird mit Hilfe der Prozedur DAD variiert. Für alle möglichen Vorzeichenkombinationen wird die Norm F berechnet und daraus das Maximum bestimmt.

f) Die real-Prozedur VNORM(VEK,N,I) berechnet eine Norm des Vektors VEC [1:N] , wobei I angibt, welche Norm zugrunde liegt.

g) Die real-Prozedur MAXEIG(A,N) berechnet den maximalen Eigenwert einer reellen symmetrischen Matrix A [1:N,1:N] .

h) Die real-Prozedur MNORM(MAT,N,I) berechnet die Norm der reellen Matrix MAT [1:N,1:N] , wobei I angibt, welche Norm zugrunde liegt.

i) Die real-Prozedur NORMAX(A,N,P) berechnet aus der Matrix A [1:N,1:N] unter Zugrundelegung der Norm P die Größe α_p.

j) Die real-Prozedur NAEMAX(A,B,N,M,P) berechnet aus der Matrix A [1:N,1:N] unter Zugrundelegung des Aufspaltungsvektors B[1:M] die Größe α_p^* . (Der Aufspaltungsvektor B [1:M] entspricht dem Vektor $u^T = (n_1,\ldots,n_r)$!).

k) Die Prozedur TRIDI bewirkt Tridiagonalisation einer symmetrischen Matrix.

l) Die Prozedur EIGVAL berechnet Eigenwerte für eine symmetrische Tridiagonalmatrix .

2. Erklärung der Eingabedaten.

Die Daten müssen sich in folgender Reihenfolge auf dem Datenstreifen befinden:

a) Die Ordnung n der vorliegenden Matrix A.

b) Die Zahl r, deren Betrag die Anzahl der Aufspaltungen für die näherungsweise Berechnung der Größe α_i angibt. (s. Kapitel III, § 2 Näherungsverfahren III). Es gilt:

$r > 0$: sowohl die exakten als auch die angenäherten Werte α_i werden berechnet.

$r < 0$: nur die angenäherten Werte α_i^* werden berechnet.

$r = 0$: nur die exakten Werte α_i werden berechnet.

c) Die Matrix A wird zeilenweise eingegeben.

d) Falls $r \neq 0$ wird der Aufspaltungsvektor $u^T = (n_1, \ldots, n_r)$ eingegeben.

3. Erklärung der Ausgabedaten.

Legen wir die in der Arbeit verwendeten Symbole zugrunde, so sind die von der Rechenmaschine ausgedruckten Bezeichnungen folgendermaßen zu verstehen:

AUFSPALTUNG : $= u^T = (n_1, \ldots, n_r)$; ALPHA : $= \alpha$

ALPHA-N1 : $= \alpha^*$ (s. Kapitel III, § 2 Näherungsverfahren III)

ALPHA-N2 : $= \alpha^{**}$ (s. Kapitel III, § 2 Näherungsverfahren I)

INDEX(NORM) gibt die zugrunde gelegte Norm an.

4. Programm

```
'begin'
  'comment' fehlerabschaetzung;
  'integer' atyp,btyp,dec;
  'procedure' invers(n,a,eps,alarm,delta);'value'n;'integer'n;'real'eps,delta;'array'a;'label'alarm;
  'begin'
    'array' b,c[1:n]; 'real' w,y; 'integer''array' z[1:n]; 'integer' i,j,k,l,p;
    delta := 1.0;
    'for' j := 1 'step' 1 'until' n 'do' z[j] := j;
    'for' i := 1 'step' 1 'until' n 'do'
    'begin'
      k := i;   y := a[i,i];   l := i-1;   p := i+1;
      'for' j := p 'step' 1 'until' n 'do'
      'begin'
        w := a[i,j];
        'if' abs(w) 'greater' abs(y) 'then'
        'begin'
```

```
            k := j;  y := w
          'end'
      'end';
      delta := delta*y;
      'if' abs y) 'less' eps 'then' 'go to' alarm;
      y := 1.0/y;
      'for' j := 1 'step' 1 'until' n 'do'
      'begin'
         c[j] := a[j,k];  a[j,k] := a[j,i];  a[j,i] := -c[j]*y;  b[j] := a[i,j] := a[i,j]*y
      'end';
      a[i,i] := y;  j := z[i];  z[i] := z[k];  z[k] := j;
      'for' k := 1'step'1'until'l,p'step'1'until'n'do'
      'for' j := 1'step'1'until'l,p'step'1'until'n'do' a[k,j] := a[k,j]-b[j]*c[k]
   'end';
   'for' i := 1 'step' 1 'until' n 'do'
   'begin'
repeat: k := z[i];
      'if' k 'equal' i 'then' 'go to'  advance;
      'for' j := 1 'step' 1 'until' n 'do'
      'begin'
         w := a[i,j];  a[i,j] := a[k,j];  a[k,j] := w
      'end';
      p := z[i];  z[i] := z[k];  z[k] := p;
      delta := -delta;  'go to' repeat;
advance:'end'
'end';
'procedure' tridi(lp,r,a,b,w,q); 'value' lp; 'integer' lp; 'array' r,a,b,w,q;
'begin'
   'integer' lp1,i,j,k;   'real' x;
   'for' i:= 1 'step' 1 'until' lp 'do' a[i]:= r[i,i];
   b[1]:= 0; lp1:= lp-1;
   'if' lp 'less' 2 'then' 'go to' tl2;
   'if' lp 'equal' 2 'then' 'go to' tl1;
   'for' k:= 2 'step' 1 'until' lp1 'do'
   'begin'
      x:= 0;
      'for' j:= k 'step' 1 'until' lp 'do' x:= x+r[j,k-1]*r[j,k-1];
      x:= sqrt(x); b[k]:= x*sign(-r[k,k-1]); x:= 1/x; w[k]:= sqrt(abs(r[k,k-1])*x+1);
      x:= (x/w[k])*sign(r[k,k-1]); r[k,k-1]:= w[k];
      'for' i:= k+1 'step' 1 'until' lp 'do' r[i,k-1]:= w[i]:= x*r[i,k-1];
      'for' j:= k 'step' 1 'until' lp 'do'
      'begin'
         q[j]:= 0;
         'for' i:= k 'step' 1 'until' j 'do' q[j]:= q[j]+r[j,i]*w[i];
         'if' j 'less' lp 'then'
            'for' i:= j+1 'step' 1 'until' lp 'do' q[j]:= q[j]+r[i,j]*w[i];
      'end';
      x:= 0;
      'for' j:= k 'step' 1 'until' lp 'do' x:= x+w[j]*q[j];
      x:= 0.5*x;
      'for' i:= k 'step' 1 'until' lp 'do' q[i]:= x*w[i]-q[i];
      'for' i:= k 'step' 1 'until' lp 'do'
      'for' j:= i 'step' 1 'until' lp 'do' r[j,i]:= r[j,i]+q[i]*w[j]+q[j]*w[i];
  'end';
```

```
    'for' i:= 1 'step' 1 'until' lp 'do'
    'begin'
       x:= a[i]; a[i]:= r[i,i]; r[i,i]:= x;
    'end';
tl1: b[lp]:= r[lp,lp1];
tl2: 'end' tridi;
'procedure' eigval(lp,e,a,b,f,w); 'value' lp; 'integer' lp; 'array' e,a,b,f,w;
'begin'
    'integer' i,j,k,n,s1,s2;   'real' bd,x;
    'real''procedure' maxi(u,v); 'value' u,v; 'real' u,v;
       maxi:= 'if' u 'greater' v 'then' u 'else' v;
    bd:= abs(a[1]);
    'for' i:= 2 'step' 1 'until' lp 'do' bd:= maxi(bd,abs(a[i])+b[i]*b[i]);
    bd:= bd*1;
    'for' i:= 1 'step' 1 'until' lp 'do'
    'begin'
       a[i]:= a[i]/bd; b[i]:= b[i]/bd; w[i]:= 1; e[i]:= -1;
    'end';
    'for' k:= 1 'step' 1 'until' lp 'do'
    'begin'
el1:  'if' (w[k]-e[k])/maxi(maxi(abs(w[k]),abs(e[k])),1_b-9) 'notgreater' 1_b-7 'then' 'go to' el6;
      x:= (w[k]+e[k])*0.5; s2:= s1:= 1; f[1]:= a[1]-x; n:= 0;
      'if' f[1] 'less' 0 'then' s1:= -1 'else' n:= 1;
      'for' i:= 2 'step' 1 'until' lp 'do'
      'begin'
         'if' b[i] 'equal' 0 'then'
            'begin'
               f[i]:= (a[i]-x)*s1; 'go to' el2;
         'end';
         'if' b[i-1] 'equal' 0 'then'
         'begin'
            f[i]:= (a[i]-x)*f[i-1]-b[i]*b[i]*s2; 'go to' el2;
         'end';
         'if' abs(f[i-1])+abs(f[i-2]) 'less' 1_b-12 'then'
         'begin'
            f[i-1]:= f[i-1]*1_b12; f[i-2]:= f[i-2]*1_b12;
         'end';
         f[i]:= (a[i]-x)*f[i-1]-b[i]*b[i]*f[i-2];
el2:     s2:= s1;
         'if' f[i] 'equal' 0 'then' 'go to' el3;
         s1:= sign(f[i]);
         'if' s1+s2 'equal' 0 'then' 'go to' el4;
el3:     n:= n+1;
el4:  'end' i-schleife;
      n:= lp-n;
      'if' n 'less' k 'then' 'go to' el5;
      'for' j:= k 'step' 1 'until' n 'do' w[j]:= x;
el5:  n:= n+1;
      'if' lp 'less' n 'then' 'go to' el1;
      'for' j:= n 'step' 1 'until' lp 'do'
        'if' x 'greater' e[j] 'then' e[j]:= x 'else' 'go to' el1;
      'go to' el1;
```

```
   ol6:'end' k-schleife;
      'for' i:= 1 'step' 1 'until' lp 'do'
      'begin'
         a[i]:= a[i]*bd; b[i]:= b[i]*bd; f[i]:= (w[i]+e[i])*bd*0.5;
      'end';
      j:= lp; k:= 1;
      'for' i:= 1 'step' 1 'until' lp 'do'
         'if' f[k] 'greater' f[j] 'then'
         'begin'
            e[i]:= f[k]; k:= k+1;
         'end'                        'else'
         'begin'
            e[i]:= f[j]; j:= j-1;
         'end';
   'end' eigval;
rep:read(atyp,btyp);
   'if' btyp'equal' 0 'then' dec:=2 'else' dec:=abs(btyp);
   'begin'
      'real' epsi,delta,deter,zaesp,nesp; 'integer' l,k,erg; 'boolean' naebe,exbe;
      'array' ain[1:atyp,1:atyp],bin[1:dec],vin[1:atyp],alpha,alphan,manor[1:3];
      'procedure'lines(p);'integer'p;
      'begin'
         'integer'pp;
         'for'pp:=1'step'1'until'p'do'nlcr;
      'end'lines;
      'procedure' output(mat,n);'integer' n;'array' mat;
      'begin'
         'integer' ll,kk,lk;
         'for' ll:=1 'step' 1 'until' n'do'
         'begin'
            kk:=0;
ol1:        'for' lk:=1 'step' 1 'until' 10 'do'
            'begin'
               kk:=kk+1;
               'if' kk'greater' n 'then'
               'begin'
                  'if'lk'equal'1'then''goto'ol3'else''goto'ol2;
               'end';
               flot(4,2,mat[ll,kk]);
            'end';
            lines(1); 'goto' ol1;
ol2:        lines(1);
ol3:        lines(2);
         'end';
      'end' output;
     'procedure' ata(aa,ab,n);'value' n;'integer' n; 'array' aa,ab;
      'begin'
         'real' s; 'integer' li,lz,ls;
         'for' lz:=1 'step' 1 'until' n 'do' 'for' ls:=1 'step' 1 'until' n 'do'
         'begin'
            s:=0;
            'for' li:=1 'step' 1 'until' n 'do' s:=s+aa[li,ls]*aa[li,lz];
            ab[lz,ls]:=s;
         'end';
      'end' ata;
```

```
      'procedure' dad(bi,n);'integer' m; 'integer' 'array' bi;
      'begin'
        'integer' ll;
        'for' ll:=m 'step' -1 'until' 1 'do'
        'begin'
          'if' bi[ll] 'equal' -1 'then' bi[ll]:=1 'else'
          'begin'
            bi[ll]:=-1; 'goto' lf;
          'end';
        'end';
lf:   'end' dad;
      'procedure'maxno(a,n,m,p,v,norm,f);'integer'n,m,p;'real'norm;'array'a,v;'real''procedure'f;
      'begin'
        'real' mmax; 'integer' lx,ly; 'array' tmax[1:m]; 'integer' 'array' vorz[1:m];
        'for' lx:=1 'step' 1 'until' m 'do'vorz[lx]:=1;
        norm:=0;
lg:     'for' lx:=1 'step' 1 'until' m 'do'
        'begin'
          tmax[lx]:=0;
          'for' ly:=1 'step' 1 'until' m 'do' tmax[lx]:=tmax[lx]+vorz[ly]*a[lx,ly];
        'end';
        mmax:=f(tmax,m,p);
        'if' mmax 'greater' norm 'then'
        'begin'
          norm:=mmax;
          'for' lx:=1 'step' 1 'until' m 'do' v[lx]:=tmax[lx];
        'end';
        dad(vorz,m);
        'if' vorz[1] 'equal' 1 'then' 'goto' lg;
      'end' maxno;
      'real''procedure' maxeig(a,n); 'integer' n; 'array' a;
      'begin'
        'array' d,sd,v,w1,w2[1:n];
        tridi(n,a,d,sd,w1,w2); eigval(n,v,d,sd,w1,w2); maxeig:= v[1];
      'end' maxeig;
      'real''procedure' vnorm(vek,n,i); 'integer' n,i; 'array' vek;
      'begin'
        'real' max; 'integer' lv; 'switch' fall:= f1,f2,f3;
        max:= 0; 'go to' fall[i];
f1:     'for' lv:= 1 'step' 1 'until' n 'do' 'if' abs(vek[lv]) 'greater' max 'then' max:= abs(vek[lv]);
        'go to' f5;
f2:     'for' lv:= 1 'step' 1 'until' n 'do' max:= max+abs(vek[lv]);
        'go to' f5;
f3:     'for' lv:= 1 'step' 1 'until' n 'do' max:= max+vek[lv]*vek[lv];
        max:= sqrt(max);
f5:     vnorm:= max;
      'end' vnorm;
```

```
      'real' 'procedure' mnorm(mat,n,i); 'integer' n,i; 'array' mat;
      'begin'
         'real' max; 'integer' ka,km; 'array' trs[1:n,1:n], mi[1:n]; 'switch' form:=g1,g2,g3,
         max:=0; 'goto' form[i];
g3:      'for' ka:=1 'step' 1 'until' n 'do'
         'begin'
            mi[ka]:=0;
            'for' km:=1 'step' 1 'until' n 'do' mi[ka]:=mi[ka]+abs(mat[ka,km]),
         'end';
         'goto' g12;
g2:      'for' ka:=1 'step' 1 'until' n 'do'
         'begin'
            mi[ka]:=0;
            'for' km:=1 'step' 1 'until' n 'do' mi[ka]:=mi[ka]+abs(mat[km,ka]);
         'end';
g12:     'for' ka:=1 'step' 1 'until' n 'do' 'if' mi[ka] 'greater' max 'then' max:=mi[ka];
         'goto' g5;
g1:      ata(mat,trs,n); max:=sqrt(maxeig(trs,n));
g5:      mnorm:=max;
      'end' mnorm;
      'real' 'procedure' normax(a,n,p); 'integer' n,p; 'array' a;
      'begin'
         'real' maxim; 'array' theta[1:n];
         maxno(a,n,n,p,theta,maxim,vnorm); normax:=1/maxim;
      'end' normax;
      'real' 'procedure' naemax(a,b,n,m,p); 'integer' n,m,p; 'array' a,b;
      'begin'
         'real' maxim; 'integer' ku,kxy,kyz,mm,nn;
         'array' ared[1:n,1:m], theta[1:n]; 'integer' 'array' bred[1:n];
         bred[1]:=0;
         'for' ku:=2 'step' 1 'until' n 'do' bred[ku]:=b[ku-1]+bred[ku-1];
         'for' ku:=1 'step' 1 'until' n 'do'
         'begin'
            mm:=b[ku]; nn:=bred[ku];
            'begin'
               'array' ateil[1:n,1:mm];
               'for' kxy:=1 'step' 1 'until' n 'do'
                  'for' kyz:=1 'step' 1 'until' mm 'do' ateil[kxy,kyz]:=a[kxy,kyz+nn];
               maxno(ateil,n,mm,p,theta,maxim,vnorm);
            'end';
            'for' kxy:=1 'step' 1 'until' n 'do' ared[kxy,ku]:=theta[kxy];
         'end';
         maxno(ared,n,m,p,theta,maxim,vnorm); naemax:=1/maxim;
      'end' naemax;
      naebe:=exbe:='false';
      'if' btyp 'equal' 0 'then' naebe:='true';
      'if' btyp 'less' 0 'then' exbe:='true';
      'for' k:=1 'step' 1 'until' atyp 'do' 'for' k:=1 'step' 1 'until' atyp 'do' read(ain[k,k]);
      'if' naebe 'then' 'goto' L1;
      'for' k:=1 'step' 1 'until' dec 'do' read(bin[k]);
L1:   lines(2); write(''eingabematrix''); lines(2); output(ain,atyp); lines(3);
      invers(atyp,ain,1,-10,sing,deter); 'go to' L11;
sing: write(''eingabematrix singulaer''); 'go to' stop;
L11:  write(''inverse matrix''); lines(2); output(ain,atyp); lines(3);
```

```
      write(''ordnung der matrix=''); absfixt(2,0,atyp); space(5);
      'if' naebe 'then' 'goto' l2;
      write(''aufspaltung:'');
      'for' i:=1 'step' 1 'until' dec 'do' absfixt(2,0,bin[i]);
l2:   lines(4);
      'for' i:=1 'step' 1 'until' 3 'do' manor[i]:=1/(manorm(ain,atyp,i));
      'if' exbe 'then' 'goto' l3;
      'for' i:=1 'step' 1 'until' 3 'do' alpha[i]:=normax(ain,atyp,i);
l3:   'if' naebe'then' 'goto' l4;
      'for' i:=1 'step' 1 'until' 3 'do' alphan[i]:=naemax(ain,bin,atyp,dec,i);
l4:   manor[2]:= manor[2]/atyp; manor[3]:= manor[3]/sqrt(atyp);
      write(''index(norm)  '');
      'if' exbe 'then' 'go to' l5;
      write(''  alpha   '');
      'if' naebe 'then' 'go to' l6;
l5:   write('' alpha-n1  '');
l6:   write('' alpha-n2''); lines(3);
      'for' i:=1 'step' 1 'until' 3 'do'
      'begin'
         space(4); absfixt(2,0,i); space(5);
         'if' exbe 'then' 'goto' l9;
         flot(4,2,alpha[i]);
         'if' naebe 'then' 'goto' l10;
l9:      flot(4,2,alphan[i]);
l10:     flot(4,2,manor[i]); lines(2);
      'end';
stop: newpage; 'go to' rep;
   'end' innerer block;
'end'
↵
```

III. Algolprogramm für Iterationsverfahren zur exakten und näherungsweisen Berechnung der Größen $\max_{\|s\|_3=1} \|As\|_3$ und $\min_{\|s\|_3=1} \|As\|_3$.

1. Bemerkungen zum Programm.

a) Die Prozedur MINMAX(A,N_0,L,N,LAMBDA,EPS) berechnet iterativ den größten bzw. den kleinsten Eigenwert der reellen symmetrischen Matrix A^TA, je nachdem ob L : = 1 oder L : = -1 gesetzt wird. A ist eine quadratische Matrix der Ordnung N. N_0 gibt an, wie viele Iterationen maximal durchgeführt werden sollen, falls $N_0 > 0$. Gilt schon bei einer geringeren Anzahl von Iterationen die folgende Beziehung:

$$(*) \quad |\lambda_{n+1} - \lambda_n| < \tfrac{1}{2} |\lambda_{n+1}| \cdot EPS ,$$

so bricht das Verfahren ab.
Wird $N_0 \leq 0$ gesetzt, so bricht das Verfahren erst ab, wenn die Beziehung $(*)$ erfüllt ist.
Bei einer Maschinengenauigkeit von n signifikanten Dezimalstellen muß gelten: $1 > EPS \geq 10^{-n+1}$
Der numerische Wert des gewünschten Eigenwertes kann unter LAMBDA abgerufen werden. LAMBDA entspricht dem in der Arbeit mit ω_κ^2 bezeichneten Wert.

b) Die Prozedur ROSI 33(A,L,N_0,LAMBDA,EPS) berechnet iterativ einen der größten bzw. einen der kleinsten Eigenwerte von A^TA, je nachdem ob L : = 1 oder L : = -1 gesetzt wird.
Sonst vergleiche man die Bemerkungen zur Prozedur MINMAX.

2. Programm

```
'begin'
'real' 'procedure' random(first);
      'boolean' first;
      'begin'
'integer' t, tdi;
'own' 'integer' i, p;
'real' rd;
'if' first 'then' 'begin' i := 2; p := -1; 'end'
'else' i := i + 1;
l:
t := abs(time);
'if' i 'notless' t/6000  'then' i := 2;
tdi := t'div'i;
'if' tdi 'equal' p 'then' 'begin' i := 2*i; 'go to' l; 'end';
random := t/i - tdi;
p := tdi;
        'end' random;
```

```
'procedure' einmax(a,n0,l,n,lambda,eps);
'integer' n0,l,n;
'real' lambda,eps;
'array' a;
'begin'
'integer' k,i,j;
'real' rho,w,laada,v2,y2,z2,w1;
'array' t[1:n],u[1:n],v[1:n],s[1:n];

rho:=1; w:=0; k:=1; w1:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
s[i]:=0; t[i]:=0; u[i]:=a[i,1]; w:=w+u[i]×u[i]; 'end';
t[1]:=1; s[1]:=1;

mk:   v2:=0; y2:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
v[i]:=0;
'for' j:=1 'step' 1 'until' n 'do'
v[i]:=v[i]+a[j,i]×u[j];
v[i]:=v[i]-w×s[i];
v2:=v2+v[i]×v[i]; 'end';
'for' i:=1 'step' 1 'until' n 'do' 'begin'
u[i]:=0;
'for' j:=1 'step' 1 'until' n 'do'
u[i]:=u[i]+a[i,j]×v[j];
y2:=y2+u[i]×u[i]; 'end';
z2:=(y2/v2)-w;
laada:=((z2+(sqrt(z2×z2+4×v2)×l))/v2)×(rho/2);
rho:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
t[i]:=t[i]+lamda×v[i]; rho:=rho+t[i]×t[i]; 'end';
rho:=sqrt(rho); w:=0; k:=k+1;
'for' i:=1 'step' 1 'until' n 'do' s[i]:=t[i]/rho;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
u[i]:=0;
'for' j:=1 'step' 1 'until' n 'do'
u[i]:=u[i]+a[i,j]×s[j];
w:=w+u[i]×u[i]; 'end';
'if' abs(w-w1) 'less' (abs(w)×eps)/2 'then' 'goto' ende;
w1:=w;
'if' k 'equal' n0 'then' 'goto' ende;
'goto' mk;
ende:  lambda:=w; 'end';
'procedure' rosi33(a,l,n,n0,lambda,eps);
'integer' l,n,n0;
'real' lambda,eps;
'array' a;

'begin'
'integer' i,j,nue,k1;
'real' s,rho,w,m,delta,lamda,w1;
'array' g[1:n],t[1:n],z[1:n],x[1:n];

nue:=1;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
```

```
s:=0;
'for' j:=1 'step' 1 'until' n 'do' s:=s+a[j,i]×a[j,i];
g[i]:=s; 'if' s 'less' g[nue] 'then'  nue:=i; 'end';
'for' i:=1 'step' 1 'until' n 'do' 'begin'
t[i]:=0; z[i]:=a[i,nue]; x[i]:=a[i,nue]; 'end';
t[nue]:=1; rho:=1; w:=g[nue]; k1:=1; w1:=w-1;
'if' nue 'less' n 'then' nue:=nue+1 'else' nue:=1;

mk:   delta:=0;
'for' i:=1 'step' 1 'until' n 'do' delta:=delta+a[i,nue]×x[i];
s:=delta-(t[nue]/rho)×w;
m:=delta-(t[nue]/rho)×g[nue];
delta:=(g[nue]-w)/m;
'if' m 'equal' 0 'then' lamda:=((-1)/rho)×(s/(g[nue]-w))
'else' lamda:=(rho/2)×(delta+1×sign(m)×sqrt(delta×delta+4×(s/m)));
rho:=sqrt(rho×rho+2×t[nue]×lamda+lamda×lamda);
t[nue]:=t[nue]+lamda;
w:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
z[i]:=(z[i]+lamda×a[i,nue]); x[i]:=z[i]/rho;
w:=w+x[i]×x[i]; 'end';
k1:=k1+1;
'if' nue 'less' n 'then' nue:=nue+1 'else' nue:=1;
'if' k1 'equal' n0 'then' 'goto' ende;
'if' abs(w-w1) 'less' (abs(w)×eps)/2 'then' 'goto' ende;
w1:=w;
'goto' mk;
ende:   lambda:=w; 'end';

'procedure'mjacobi(a,l,n,eps);
'value'n,eps;
'integer'n;
'real'eps;
'array'a,l;

'begin'
'integer'i,j,k,m,p,q,ind,ss;
'real'sum,norm,norm1,s,con,con1,app,aqq,c,sint,cost;
'integer''array'it[1:n];

s:=4×n;
ind:=0;   sum:=0;
k:=(n×(n+1))/2;

'for'i:=1'step'1'until'n'do'
it[i]:=(i×(i+1))/2;

'for'j:=2'step'1'until'k'do'
sum:=sum+a[j]×a[j];

'for'j:=2'step'1'until'n'do''begin'
```

```
i:=it[j];
sum:=sum-a[i]*a[i];
'end';

norm:=sqrt(sum);
norm1:=norm*eps/n;

2: s:=s*s;
norm:=norm/s;
9: q:=2;
7: p:=1;
5: j:=(q*(q-1))/2;
con:=-a[j+p];
'if'abs(con)'notgreater'norm'then''goto'3;
ind:=1;
k:=(p*(p+1))/2;
app:=a[k];    k:=(q*(q+1))/2;
aqq:=a[k];
con1:=0.5*(app-aqq);
s:=1;
'if'con1'less'0'then'ss:=-1;

c:=ss*con/sqrt(con*con+con1*con1);
sint:=c/sqrt(2*(1+sqrt(1-c*c)));
cost:=sqrt(1-sint*sint);

m:=(q*(q-1))/2;
'if'p+1'equal'q'then''goto'11;
'for'i:=p+1'step'1'until'q-1'do''begin'
j:=(i*(i-1))/2;
c:=a[j+p];    k:=m+i;
a[j+p]:=c*cost-a[k]*sint;
a[k]:=c*sint+a[k]*cost;
'end';

11: 'if'p'equal'1'then''goto'12;
j:=(p*(p-1))/2;
'for'i:=1'step'1'until'p-1'do''begin'
c:=a[j+i];    k:=m+i;
a[j+i]:=c*cost-a[k]*sint;
a[k]:=c*sint+a[k]*cost;
'end';

12: 'if'q'equal'n'then''goto'13;
'for'i:=q+1'step'1'until'n'do''begin'
j:=(i*(i-1))/2;
c:=a[j+p];    k:=j+q;
a[j+p]:=c*cost-a[k]*sint;
a[k]:=c*sint+a[k]*cost;
'end';

10: c:=2*sint*cost;
j:=(p*(p+1))/2;    m:=(q*(q+1))/2;
```

```
a[j]:=app*cost*cost+aqq*sint*sint+c*con;
a[m]:=app*sint*sint+aqq*cost*cost-c*con;
k:=(q*(q-1))/2+p;
a[k]:=0;


3: 'if'p'equal'q-1'then''goto'4;
p:=p+1;
'goto'5;
4: 'if'q'equal'n'then''goto'6;
q:=q+1;
'goto'7;
6: 'if'ind'equal'0'then''goto'8;
ind:=0;   'goto'9;
8: 'if'norm'greater'norm1'then''goto'2;
'for'i:=1'step'1'until'n'do''begin'
j:=it[i];
l[i]:=a[j];
'end';

'end';

'begin'
'integer' n,n0,l,i,j,m,m0,mm;
'real' b,eps,lambda;
'begin'
read(n); mm:=(n*(n+1))/2;
'begin'
'array' a[1:n,1:n],bb[1:mm],ll[1:n];
b:=random('true');
'for' i:=1 'step' 1 'until' 50 'do' b:=random('false');
read(n0,m,m0);
'for' i:=1 'step' 1 'until' n 'do' 'begin'
'for' j:=1 'step' 1 'until' n 'do' 'begin'
b:=(random('false')-0.5)*2*m; a[i,j]:=b; 'end'; 'end';
l:=1; nlcr; eps:=1.-10; lambda:=0;
write(''maximaler eigenwert mit minmax''); nlcr;
minmax(a,n0,l,n, lambda,eps); print(lambda);
write(''maximaler eigenwert mit rosi33''); nlcr;
rosi33(a,l,n,n0,lambda,eps); print(lambda);
l:=-1; nlcr;
write(''minimaler eigenwert mit minmax''); nlcr;
minmax(a,n0,l,n,lambda,eps); print(lambda);
write(''minimaler eigenwert mit rosi33''); nlcr;
rosi33(a,l,n,n0,lambda,eps); print(lambda);
```

```
write('' wurzel der eigenwerte von ata''); nlcr;
'for' m:=1 'step' 1 'until' n 'do' 'begin'
'for' i:=m 'step' 1 'until' n 'do' 'begin'
ll[i]:=0;
'for' j:=1 'step' 1 'until' n 'do' ll[i]:=ll[i]+a[m,j]×a[i,j]; 'end';
'for' i:=m 'step' 1 'until' n 'do' a[m,i]:=ll[i]; 'end';
l:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
l:=l+i-1;
'for' j:=1 'step' 1 'until' i 'do' bb[l+j]:=a[j,i]; 'end';
'for' i:=1 'step' 1 'until' n 'do' ll[i]:=0;
eps:=1₁₀-10;
mjacobi(bb,ll,n,eps);
j:=0;
'for' i:=1 'step' 1 'until' n 'do' 'begin'
j:=j+1; 'if' j 'equal' 5 'then' 'begin' nlcr; j:=1; 'end';
flot(13,3,sqrt(ll[i])); 'end';
'end'; 'end'; 'end'; 'end';⏎
```

IV. Algolprogramm für ein Iterationsverfahren zur näherungsweisen Berechnung der Größe $\min_{\|s\|_2=1} \|A^T s\|_3 = \alpha_3$.

1. Bemerkung zu den Prozeduren im Programm.
   Neben der Bibliotheksprozedur RANDOM werden nur einige Hilfsprozeduren verwendet.
2. Erklärung der Eingabedaten.

a) M_o : Angabe von Zwischenergebnissen alle M_o Iterationen.
b) N_o : Zahl der gewünschten Iterationen.
c) V : Ordnung der Matrix A.
d) SCR1 : Definitionsbereich für die Elemente von A.
   Es gilt: $-SCR1 \leq a_{ik} \leq +SCR1$

3. Erklärung der Ausgabedaten.
   Alle M_o Iterationen wird der Wert ω_κ ausgedruckt.
4. Programm

```
'begin'
'integer'  v,scr1,m0,n0,m,n,v1,p;
'real' 'procedure' random(first);
      'boolean' first;
      'begin'
'integer' t, tdi;
'own' 'integer' i, p;
'real' rd;
'if' first 'then' 'begin' i := 2; p := -1; 'end'
'else' i := i + 1;
L:
t := abs(time);
'if' i 'notless' t/6000  'then' i := 2;
tdi := t'div'i;
'if' tdi 'equal' p 'then' 'begin' i := 2*i; 'go to' L; 'end';
random := t/i - tdi;
p := tdi;
      'end' random;
'procedure'  pdelta(a,x,v,delta,vp);
'array'  a,x,delta;
'integer'  v,vp;
'begin'
'integer'  i,j;
i:=vp;
delta[i]:=0;
'for'  j:=1 'step' 1 'until'  v 'do'  delta[i]:=delta[i]+x[j]*a[j,i];
'end';
'procedure'  ps(delta,omega,t,v,s,vp);
'array'  delta,t,s;
'real'  omega;
'integer'  v,vp;
'begin'
'integer'  i;
i:=vp;
'if'  t[i]  'greater'  0 'then'   s[i]:=delta[i]-omega;
```

```
'if' t[i] 'less' 0 'then' s[i]:=delta[i]+omega;
'if' t[i] 'equal' 0 'then' s[i]:=delta[i]+omega*sign(delta[i]);
'end';

'procedure' plambda(ro,gamma,omega,delta,s,t,v,lambda,vp);
'array' gamma,delta,s,t,lambda;
'real' ro,omega;
'integer' v,vp;
'begin'
'integer' i;
'real' res1,res2,res3;
i:=vp;
res1:=(gamma[i]-omega)/(delta[i]-gamma[i]);
res2:=(gamma[i]-omega)/(delta[i]+gamma[i]);
res3:=(gamma[i]-omega)/(delta[i]+gamma[i]*sign(delta[i]));

'if' t[i] 'greater' 0 'then' lambda[i]:=ro*0.5*(res1+sqrt(res1*res1+(4*s[i])/(delta[i]-gamma[i])));
'if' t[i] 'less' 0 'then' lambda[i]:=ro*0.5*(res2-sqrt(res2*res2+(4*s[i])/(delta[i]+gamma[i])));
'if' t[i] 'equal' 0 'then' lambda[i]:=ro*0.5*(res3-sign(delta[i])*sqrt(res3*res3+4*s[i]/
(delta[i]+gamma[i]*sign(delta[i]))));
'end';

'procedure' ptz(t,v,lambda,z,a,ro,x,omega,vp);
'array' x,t,lambda,z,a;
'integer' v,vp;
'begin'
'integer' i,j;
i:=vp;
'if' t[i]*lambda[i] 'notless' -t[i]*t[i] 'then'
'begin'
t[i]:=t[i]+lambda[i];
ro:=ro-abs(t[i])+abs(t[i]+lambda[i]);
'for' j:=1 'step' 1 'until' v 'do' z[j]:=z[j]+lambda[i]*a[j,i];
'end''else'
'begin'
t[i]:=0;
ro:=ro-abs(t[i]);
'for' j:=1 'step' 1 'until' v 'do' z[j]:=z[j]-t[i]*a[j,i];
'end';
omega:=0;
'for' i:=1 'step' 1 'until' v 'do' 'begin' x[i]:=z[i]/ro;
omega:=omega+x[i]*x[i];
'end';
'end';

newpage;
write(''berechnung von rho32(a) '');
nlcr;
nlcr;
m0:=hand(1);
n0:=hand(2);
v:=hand(3);
scr1:=hand(4);
```

```
'begin'
'array'a[1:v,1:v],t,gamma,z,x,delta,s,lambda[1:v];
'real' ro,omega;
'integer' i,j;
t[1]:=random('true');
'for' i:=1 'step' 1 'until' v'do' 'begin' gamma[i]:=0;
'for' j:=1 'step' 1 'until' v 'do' 'begin' a[j,i]:=2*(random('false')-0.5)*scr1;
gamma[i]:=gamma[i]+a[j,i]*a[j,i];
'end';
'end';
v1:=1;
'for' i:=1 'step' 1 'until' v 'do' 'begin' 'if' gamma[i] 'notgreater' gamma[v1] 'then' v1:=i; 'end';

'for' i:=1 'step' 1 'until' v 'do' 'begin' 'if' i 'equal' v1 'then' t[i]:=1 'else't[i]:=0;
x[i]:=z[i]:=a[i,v1]; 'end';
ro:=1;
omega:=gamma[v1];
p:=m:=n:=1;
rp:
pdelta(a,x,v,delta,p);
ps(delta,omega,t,v,s,p);
plambda(ro,gamma,omega,delta,s,t,v,lambda,p);
ptz(t,v,lambda,z,a,ro,x,omega,p);
p:=p+1;
'if' p 'equal' v+1 'then' p:=1;
m:=m+1;
n:=n+1;
'if' m 'equal' m0 'then' 'begin'
 space(10); write(''sqrt(omega)'');print(sqrt(omega));
'goto' finis;
'end';
'if' m 'equal' m0 'then' 'begin' m:=0;

 space(15); ; print(sqrt(omega));
'goto' rp;
'end'
'else' 'goto' rp;
finis: 'end';'end';⊕
```

V. Algolprogramm für die näherungsweise Berechnung der Zeilen- und Spaltennorm einer Matrix.

1. Bemerkungen zu den Prozeduren im Programm.

a) Die Prozedur ORTHOGONAL(IT,N,A,INV,I) berechnet eine grobe Näherungsmatrix für die inverse Matrix INV = A^{-1}.

IT : Anzahl der Nachiterationen zur Verbesserung von INV. Für dieses Programm ist IT gleich Null zu wählen.

N : Ordnung der Matrix A.

A : Quadratische Matrix

INV: invertierte Matrix

I : Kontrollmatrix I = INV·A

Nach Abschluß der Prozedur wird ORTHOGONAL der Betrag der Determinante von A zugewiesen.

b) Die Prozedur NORMA(A,I,NORM) berechnet die Norm $\sum_{i=1}^{n} |a_i|$ eines Vektors a.

c) Die Prozedur EXDIT(A,U,NNULL) bewirkt die Ausgabe der Ergebnisse.

2. Erklärung der Eingabedaten.

a) M : Ordnung der Matrix A.

b) N_o: Anzahl der Iterationen.

c) LIMIT : Definitionsbereich für die Elemente von A.
Es gilt: -LIMIT ≤ a_{ik} ≤ +LIMIT

d) IT : Zahl der gewünschten Nachiterationen zur Verbesserung der inversen Matrix von A.

3. Erklärung der Ausgabedaten.

Nach jeder Iteration wird der Wert $\left(a_{\nu_\kappa}^{-} u_\kappa\right) u_\kappa$ ausgedruckt.

4. Programm

```
  'begin'
'integer'  m,n0;

'real''procedure'     orthogonal(it,n,a,inv,i);
'value'               n;
'integer'             n,it;
'array'               a,inv,i;
  'begin'
  'real'              norm,det;
  'integer'           j,k,l;
  'array'             c[1:n,1:n];

  'real''procedure'   skalar(r,s);
  'value'             r,s;
  'integer'           r,s;
```

```
    'begin'
    'real'          sum;
    'integer'       i;
    sum:=0;
    'for' i:=1 'step' 1 'until' n 'do'
       'begin'
       sum:=sum+inv[r,i]×inv[s,i];
       'end';
    skalar:=sum;
    'end';

'procedure'         mult(a,b,c);
'array'             a,b,c;
    'begin'
    'real'          s;
    'integer'       i,j,k;
    'for' i:=1 'step' 1 'until' n 'do'
       'begin'
       'for' j:=1 'step' 1 'until' n 'do'
          'begin'
          s:=0;
          'for' k:=1 'step' 1 'until' n 'do'
             'begin'
             s:=s+a[i,k]×b[k,j];
             'end';
          c[i,j]:=s;
          'end';
       'end';
    'end';

det:=0;
'for' j:=1 'step' 1 'until' n 'do'
    'begin'
    'for' k:=1 'step' 1 'until' n 'do'
       'begin'
       inv[j,k]:=a[j,k];
       c[j,k]:=0;
       'end';
    c[j,k]:=1;
    'end';
'for' j:=1 'step' 1 'until' n 'do'
    'begin'
    norm:=sqrt(skalar(j,j));
    'if' norm 'less' 1₁₀-9 'then' 'go to' o1;
    'for' k:=1 'step' 1 'until' n 'do'
       'begin'
       inv[j,k]:=inv[j,k]/norm;
       c[j,k]:=c[j,k]/norm;
       'end';
    'if' j 'equal' n 'then' 'go to' o2;
    'for' k:=j+1 'step' 1 'until' n 'do'
       'begin'
       norm:=skalar(j,k);
       'for' l:=1 'step' 1 'until' n 'do'
```

```
            'begin'
            inv[k,l]:=inv[k,l]-norm*inv[j,l];
            c[k,l]:=c[k,l]-norm*c[j,l];
            'end';
          'end';
02:   'end';
   det:=1;
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     det:=det/c[k,k];
     'end';
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     'for' l:=1 'step' 1 'until' n 'do'
       'begin'
       i[l,k]:=inv[k,l];
       'end';
     'end';
   mult(i,c,inv);
   j:=0;
03:mult(a,inv,i);
   'if' j 'equal' it 'then' 'go to' 01;
   j:=j+1;
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     'for' l:=1 'step' 1 'until' n 'do'
       'begin'
       i[k,l]:=-i[k,l];
       'end';
     i[k,k]:=2+i[k,k];
     'end';
   mult(inv,i,c);
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     i[k,k]:=i[k,k]-1;
     'end';
   mult(i,i,inv);
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     inv[k,k]:=1+inv[k,k];
     'end';
   mult(c,inv,i);
   'for' k:=1 'step' 1 'until' n 'do'
     'begin'
     'for' l:=1 'step' 1 'until' n 'do'
       'begin'
       inv[k,l]:=i[k,l];
       'end';
     'end';
   'go to' 03;
01:ortnogonal:=det;
   'end';
```

```
'real' 'procedure' random(first);
      'boolean' first;
      'begin'
'integer' t, tdi;
'own' 'integer' i, p;
'real' rd;
'if' first 'then' 'begin' i := 2; p := -1; 'end'
'else' i := i + 1;
L:
t := abs(time);
'if' i 'notless' t/6000  'then' i := 2;
tdi := t'div'i;
'if' tdi 'equal' p 'then' 'begin' i := 2*i; 'go to' L; 'end';
random := t/i - tdi;
p := tdi;
         'end' random;
'procedure'  norma(a,i,norm);
'array'  a;
'integer'  i;
'real' norm;

'begin'
'integer'  k;
norm:=0;
'for' k:=1 'step' 1 'until'  m'do'
'begin'
norm:=norm+abs(a[i,k]);
'end';
'end';
'procedure'  exdit(a,u,nnull);
'array'  a,u;
'real'  nnull;
'begin'
'real'  betrag,betragu;
'integer'  k;

betrag:=0;
betragu:=0;
'for' k:=1 'step' 1 'until'  m 'do'
'begin'
betrag:=betrag+a[k,nnull]*u[k];
betragu:=betragu+abs(u[k]);

'end';
print(betrag,betragu);
'end';
read(m,m0);
'begin'
'array'  a[1:m,1:m],gamma[1:m],u[1:m],inv[1:m,1:m],iv[1:m,1:m];
'real'  lambda,limit,eps,delta,max,result;
'integer'  p,m,it,i,j,nnull,k;
read(limit,it);
ncr;
delta:=random('true');
```

```
'for' i:=1 'step' 1 'until'  n 'do'
'begin'
'for' j:=1 'step' 1 'until' n 'do'  a[i,j]:=2×limit×(random('false')-0.5);'end';
outarray(drum,200000,a);
hold(a);
inarray(drum,200000,a);
nlcr;
hold(a);
delta:=orthogonal(it,m,a,inv,iv);
outarray(drum,300000,inv);
hold(inv);
inarray(drum,300000,a);
hold(a);
max:=0;
'for' i:=1 'step' 1 'until'  n 'do'
'begin'norma(a,i,result);

'if'  result 'greater'  max 'then'  'begin'  max:=result;nnull :=i;
'for' j:=1 'step' 1 'until' n 'do'  u[j]:=a[i,j];  'end';
'end';
outarray(drum,400000,a);
hold(a);
inarray(drum,200000,a);
hold(a);
'for' i:=1 'step' 1 'until'  n 'do'
'begin'
gamma[i]:=0;
'for' j:=1 'step' 1 'until'  n 'do'
'begin'
gamma[i]:=gamma[i]+a[j,i]×a[j,i];
'end';
'end';
p:=0;
m:=0;
rep:  p:=p+1;
'if'  p 'equal'  nnull 'then'   p:=p+1;
'if'  p 'equal'  n+1  'then'  p:=p-n;
k:=p;
Lambda:=0;
nlcr;
'for' i:=1 'step' 1 'until'  n 'do'
'begin'
'if'  i 'equal'  nnull  'then'  'goto'   just;
  Lambda:=Lambda+u[i]×a[i,k]/gamma[k];
just:
'end';
'for' i:=1 'step' 1'until'  n 'do'  u[i]:=u[i]-Lambda×a[i,k];
m:=m+1;
edit(a,u,nnull);
'if'  m  'equal'  n0   'then'  'goto'  stop 'else'  'goto'  rep;
'end';
stop:   'end';
```

Literaturverzeichnis

[1] Collatz — Zur Fehlerabschätzung bei linearen Gleichungssystemen
ZAMM 34 (1954) S. 71-72

[2] Dück — Zur Abschätzung der Fortpflanzung der Datenfehler bei linearen Gleichungssystemen nach der Formel von Wittmeyer-Collatz
ZAMM 41 (1961) S. 220-222

[3] Faddajew/Faddajewa — Numerische Methoden der linearen Algebra
S. 149-159 ; S. 522 § 74
Oldenburg Verlag München, Wien 1964

[4] Fox — An Introduction to numerical linear algebra
S. 136-142

[5] Schmetterer — Einführung in die mathematische Statistik
S. 93
Wien, New Jork 1966

[6] Turing — Rounding - off Errors in Matrix Processes
S. 287-308
Quart. J. Mech. Appl. Math. 1 (1948)

[7] Wilkinson — The algebraic eigen valueproblem
S. 189-191
Oxford 1965

[8] Wittmeyer — Einfluß der Änderung einer Matrix auf die Lösung des zugehörigen Gleichungssystems, sowie auf die charakteristischen Zahlen und Eigenvektoren
Dissertation, Darmstadt 1934

[9] Zurmühl — Matrizen und ihre technischen Anwendungen § 16
Springer-Verlag
Berlin/Göttingen/Heidelberg 1964
4. Auflage

L E B E N S L A U F

Am 9. März 1943 wurde ich in Berlin-Charlottenburg als Sohn des Regierungsrats Dr. Friedrich Mayer und seiner Ehefrau Erna, geb. Popp geboren. Bereits im Herbst 1943 siedelten wir im Rahmen der Evakuierung Berlins nach Bayreuth über. Nach vier Jahren Volksschule besuchte ich neun Jahre lang das Humanistische Gymnasium Christian-Ernestinum in Bayreuth und schloß im Juli 1962 mit dem Reifezeugnis ab. Im Jahre 1954 starb mein Vater, der zuletzt am Verwaltungsgericht in Bayreuth tätig war. Im Wintersemester 1962/63 bis einschließlich Sommersemester 1968 studierte ich an der Julius-Maximilian-Universität in Würzburg Mathematik/Physik und unterzog mich in dieser Zeit folgenden Prüfungen :

Sommer 1964	:	Vorexamen in Mathematik/Physik
Mai 1965	:	Vordiplom in Mathematik
3. Mai 1968	:	Hauptdiplom in Mathematik

Seit dem Abschluß meines Studiums arbeite ich im Rahmen eines Forschungsauftrages der Deutschen Forschungs-Gemeinschaft unter der Leitung von Prof. Dr. P.O. Runck an meiner Dissertation.

VERVIELFÄLTIGUNG

E. Schmitt u. M. Meyer
Würzburg
Alte Kasernstraße 7
Tel. 591374

VERVIELFÄLTIGUNG

E. Schmitt u. M. Meyer
Würzburg
Alte Kasernstraße 7
Tel. 591374